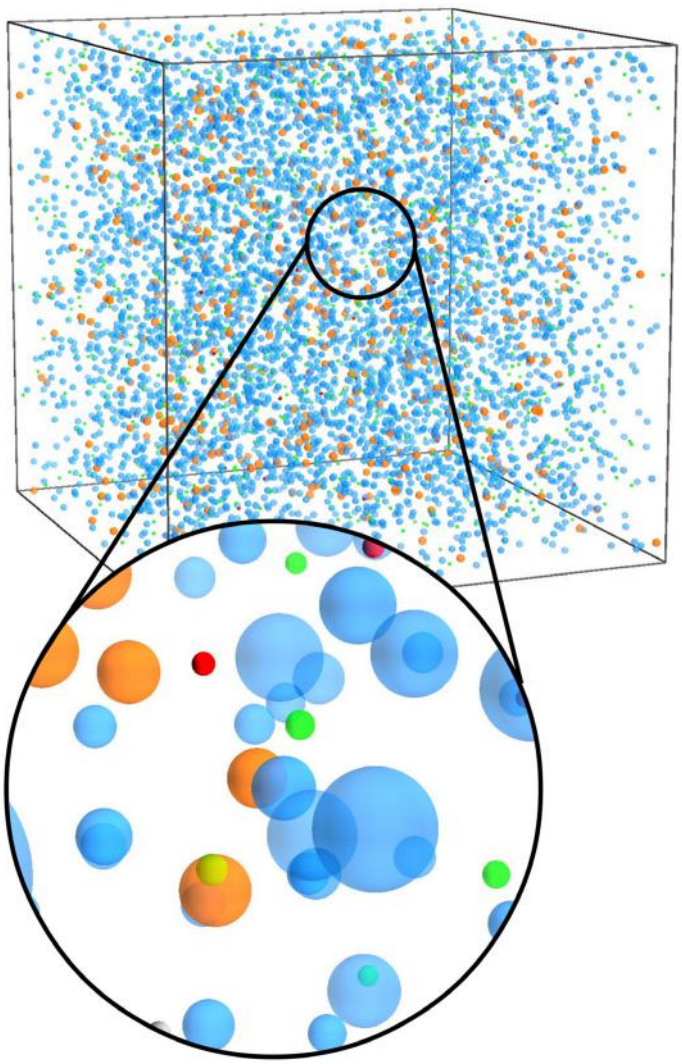
COMPUTATIONAL SYSTEMS BIOLOGY OF PATHOGEN-HOST INTERACTIONS

EDITED BY: Saliha Durmuş, Tunahan Çakır and Reinhard Guthke
PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-821-4
DOI 10.3389/978-2-88919-821-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

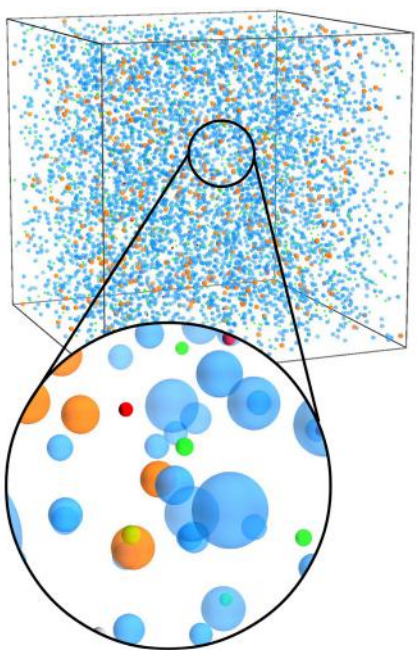
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

COMPUTATIONAL SYSTEMS BIOLOGY OF PATHOGEN-HOST INTERACTIONS

Topic Editors:
Saliha Durmuş, Gebze Technical University, Turkey
Tunahan Çakır, Gebze Technical University, Turkey
Reinhard Guthke, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knoell-Institute, Germany



Visualization of the 3D cuboid environment of the agent-based model that corresponds to 1µl of the whole-blood infection assay, containing 5000 polymorphonuclear neutrophils, 500 monocytes, and 1000 *Candida albicans* cells.

Image taken from: Lehnert T, Timme S, Pollmächer J, Hünniger K, Kurzai O and Figge MT (2015) Bottom-up modeling approach for the quantitative estimation of parameters in pathogen-host interactions. Front. Microbiol. 6:608. doi: 10.3389/fmicb.2015.00608

A thorough understanding of pathogenic microorganisms and their interactions with host organisms is crucial to prevent infectious threats due to the fact that Pathogen-Host Interactions (PHIs) have critical roles in initiating and sustaining infections. Therefore, the analysis of infection mechanisms through PHIs is indispensable to identify diagnostic biomarkers and next-generation drug targets and then to develop strategic novel solutions against drug-resistance and for personalized therapy. Traditional approaches are limited in capturing mechanisms of infection since they investigate hosts or pathogens individually. On the other hand, the systems biology approach focuses on the whole PHI system, and is more promising in capturing infection mechanisms. Here, we bring together studies on the below listed sections to present the current picture of the research on Computational Systems Biology of Pathogen-Host Interactions:

- Computational Inference of PHI Networks using Omics Data
- Computational Prediction of PHIs
- Text Mining of PHI Data from the Literature
- Mathematical Modeling and Bioinformatic Analysis of PHIs

Computational Inference of PHI Networks using Omics Data

Gene regulatory, metabolic and protein-protein networks of PHI systems are crucial for a thorough understanding of infection mechanisms. Great advances in molecular biology and biotechnology have allowed the production of related omics data experimentally. Many computational methods are emerging to infer molecular interaction networks of PHI systems from the corresponding omics data.

Computational Prediction of PHIs

Due to the lack of experimentally-found PHI data, many computational methods have been developed for the prediction of pathogen-host protein-protein interactions. Despite being emerging, currently available experimental PHI data are far from complete for a systems view of infection mechanisms through PHIs. Therefore, computational methods are the main tools to predict new PHIs. To this end, the development of new computational methods is of great interest.

Text Mining of PHI Data from Literature

Despite the recent development of many PHI-specific databases, most data relevant to PHIs are still buried in the biomedical literature, which demands for the use of text mining techniques to unravel PHIs hidden in the literature. Only some rare efforts have been performed to achieve this aim. Therefore, the development of novel text mining methods specific for PHI data retrieval is of key importance for efficient use of the available literature.

Mathematical Modeling and Bioinformatic Analysis of PHIs

After the reconstruction of PHI networks experimentally and/or computationally, their mathematical modeling and detailed computational analysis is required using bioinformatics tools to get insights on infection mechanisms. Bioinformatics methods are increasingly applied to analyze the increasing amount of experimentally-found and computationally-predicted PHI data.

Acknowledgements

We, editors of this e-book, acknowledge Emrah Nikerel (Yeditepe University, Turkey) and Arzucan Özgür (Boğaziçi University, Turkey) for their contributions during the initiation of the Research Topic.

Citation: Durmuş, S., Çakır, T., Guthke, R., eds. (2016). Computational Systems Biology of Pathogen-Host Interactions. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-821-4

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 04 February 2016
doi: 10.3389/fmicb.2016.00021

Editorial: Computational Systems Biology of Pathogen-Host Interactions

Saliha Durmuş[1], Tunahan Çakır[1] and Reinhard Guthke[2]*

[1] Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey,
[2] Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany

Keywords: pathogen-host interaction, computational systems biology, bioinformatics, omics data, network inference, text mining, constraint-based modeling, image-based systems biology

The Editorial on the Research Topic

Computational Systems Biology of Pathogen-Host Interactions

Pathogen-Host Interactions (PHIs) play a significant role in the mechanisms of infections. Therefore, the investigation of infection mechanisms through PHIs is a crucial step to develop novel and more effective solutions against drug-resistance and for personalized therapy. To this aim, systems biology approach considers the whole PHI system instead of focusing hosts or pathogens individually. Computational modeling and analysis has a vital place within the whole systems biology workflow (Cyclic operation of experimental and modeling work). Multi-scale modeling provides the holistic view needed in the investigation of pathogen-host molecular interactions. However, it is usually very difficult to identify the model structure and parameters for complex multi-scale models. On the other hand, focused modeling types require more stringent and advanced feature selection approaches.

This research topic aims to provide examples from the current picture of the research on computational systems biology of PHIs. The papers included here review recent studies or present original research on computational inference of PHI networks, computational prediction of PHIs, text mining of PHI data from the literature, and mathematical modeling and computational analysis of PHI networks. This research topic presents three review papers, 10 original research articles, and one technology report.

Opening this research topic, we provide a comprehensive review of the studies on computational systems biology of PHIs (Durmuş et al.). We focus on the computational methods for the inference of molecular interaction networks of PHI systems, bioinformatic analysis of PHI networks, the Web-based PHI databases, and text-mining efforts to extract PHI data hidden in the literature. In this sense, this review provides a systems perspective on which the other articles covered in this research topic are based.

PHI NETWORK INFERENCE USING OMICS DATA

Schulze et al. deal with the challenge of the inference of inter-species gene regulatory networks from dual transcriptomic data. They use an extended version of NetGenerator, an ordinary differential equations (ODEs)-based tool for network inference that predicts gene regulatory networks from gene expression time series data (Guthke et al., 2005; Tierney et al., 2012).

OPEN ACCESS

Edited by:
Rustam Aminov, Technical University of Denmark, Denmark

Reviewed by:
Chuang Ma, Northwest Agricultural and Forestry University, China

****Correspondence:***
Saliha Durmuş
salihadurmus@gtu.edu.tr

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *10 December 2015*
Accepted: *11 January 2016*
Published: *04 February 2016*

Citation:
Durmuş S, Çakır T and Guthke R (2016) Editorial: Computational Systems Biology of Pathogen-Host Interactions. Front. Microbiol. 7:21. doi: 10.3389/fmicb.2016.00021

Budak et al. use a temporal phosphoproteomic dataset of Salmonella-infected human cells (Rogers et al., 2011) to reconstruct the temporal signaling network of the human host by integrating protein-protein interaction (PPI) and the phosphoproteomic data. The Prize-collecting Steiner Forest approach and the Integer Linear Programming based edge inference approach are employed. The complementary use of both methods leads to a network which conserves the information about temporality, direction of interactions, while revealing the hidden entities in the signaling.

COMPUTATIONAL PREDICTION OF PHIs

Despite the recent advances, the experimentally-found PHI data are still scarce and the computational prediction is a valuable source of PHI data currently. The computational prediction primarily exploits sequence information, protein structure and known interactions. Machine learning techniques are used when there are sufficient known interactions available to be used as training data. On the opposite case, transfer and multitask learning methods are preferred. Nourani et al. provide an overview of these approaches for predicting PHIs.

Experimentally verified data on fungi-host interactions are rare in the literature and in the PHI databases. Remmele et al. reconstruct large-scale PHI networks for the fungal pathogens *Aspergillus fumigatus* and *Candida albicans* and their human and mouse hosts. A computational PHI prediction method based on protein orthology, PPI data as well as data on gene functions and cellular localization was developed and used.

TEXT MINING OF PHI DATA

The emergence of large-scale experimental PHI data has led to the development of PHI databases such as VirusMentha (Calderone et al., 2015), VirhostNet (Guirimand et al., 2015), PATRIC (Wattam et al., 2014), HPIDB (Kumar and Nanduri, 2010), and PHISTO (Durmuş Tekir et al., 2013). Nevertheless, most data regarding PHIs are still buried in the articles and they have not been stored in databases. Karadeniz et al. extend text mining tool SciMiner, originally developed for extracting intra-species molecular interactions, for inter-species PHIs. They use SciMiner to extract host-*Brucella* gene-gene interactions, which are further analyzed by ontology modeling.

MATHEMATICAL MODELING AND BIOINFORMATIC ANALYSIS OF PHIs

Few examples of constraint-based PHI models are currently available in the literature. However, there is a lack of definite description of the methodology required for the functional integration of genome scale metabolic models in order to generate PHI models. Jamshidi and Raghunathan outline a systematic procedure to produce functional PHI models, highlighting steps which require debugging and iterative revisions in order to successfully build a functional model. The construction of such models will enable the exploration of PHIs by leveraging the growing wealth of omics data in order to better understand mechanisms of infection and identify novel therapeutic strategies.

Dühring et al. describe the cross-talk between the fungal pathogen *C. albicans* and the human innate immune system. They review computational systems biology approaches to model and investigate these complex interactions with a special focus on fungal immune evasion and game-theoretical and agent-based models.

Nguyen et al. use ODEs to represent the basic interactions between Ebola virus and wild-type Vero cells, i.e., epithelial cells of green monkeys, *in vitro*. The parameters in viral kinetics are estimated leading to a first mathematical model for Ebola virus infection.

Dix et al. examine the transcriptional footprint of the host in response to the bacterial pathogens *Staphylococcus aureus* and *Escherichia coli* and the fungal pathogens *C. albicans* and *A. fumigatus* in a human whole-blood model. Expression data are exploited to build a random forest classifier to classify if a sample contains a bacterial, fungal or mock-infection.

Sinclair et al. develop a method combining *in silico* prediction of bacterial nucleomodulins, i.e., proteins targeted to the host cell nucleus, and iTRAQ protein profiling (a mass spectrometric technique where two protein expression profiles are compared) to identify potential bacterial-derived nuclear-translocated proteins that could impact transcriptional programming in host cells. This approach was applied to intracellular bacteria such as *Anaplasma phagocytophilum*, *Mycobacterium tuberculosis*, and *Chlamydia trachomatis*.

Finally, the research topic includes articles focusing on image-based systems biology of PHIs. While advances in omics techniques drive the progress of system biology on molecular level, there is also a significant progress on the cellular level based spatio-temporal data, e.g., microscopy images. Lehnert et al. apply non-spatial state-based modeling and agent-based modeling approaches to simulate an experimental assay for *C. albicans* infection of human blood. They predict cell migration parameters in 3D space where monocytes, granulocytes, and *C. albicans* cells are treated as migrating and interacting agents. Pollmächer and Figge implement a hybrid agent-based spatio-temporal modeling approach for *A. fumigatus* infection in human alveoli to decipher chemokine properties. They found by model simulations that the ratio of chemokine secretion rate to the diffusion coefficient is the main indicator for the success of pathogen detection by alveolar macrophages. Kraibooj et al. suggest a novel image analysis algorithm for the automated quantification of the phagocytosis of two wild type *A. fumigatus* strains. The strains were compared in terms of the phagocytosis process when the fungal conidia interact with alveolar macrophages.

The computational modeling of PHI networks of interacting genes, transcripts, proteins, and metabolites is crucial to enlighten the molecular mechanisms of infection. The experimental detection of levels of biomolecules via omics approaches as well as the detection of PHIs via high-throughput experiments started to generate comprehensive datasets. The modeling of the large-scale data will not only elucidate the

mechanisms of infection, but will help in the discovery of biomarkers for novel diagnostic tools and of therapeutic drug targets through identification of essential molecules for the pathogen. Despite the recent efforts, the use of systems biology approaches to investigate PHI systems is still in its infancy, mostly because of data scarcity (Durmuş et al.). Ongoing studies in the field will certainly produce more large-scale PHI data in the near future. Heterogeneous data sets (clinical, microbiological, chemical, molecular on different levels such as SNPs, transcriptome, proteome, FACS, microscopic, mass spectrometric, etc.) will be integrated. More complete PHI models will allow the integration of omics-based and image-based systems biology of infection and will pioneer more complex multi-scale models with different scale in space (from molecules/cells/tissues to organism/population) and time (from seconds to month). These more complex models will improve the PHI-based solutions to infectious diseases.

AUTHOR CONTRIBUTIONS

SD conceived the content and drafted the manuscript; TC and RG conceived the content and revised the manuscript.

ACKNOWLEDGMENTS

TC was supported by the Turkish Academy of Sciences - Outstanding Young Scientists Award Program (TÜBA-GEBİP). RG was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre/Transregio 124 FungiNet (subprojects B3 and INF).

REFERENCES

Calderone, A., Licata, L., and Cesareni, G. (2015). VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res.* 43, D588–D592. doi: 10.1093/nar/gku830

Durmuş Tekir, S., Çakir, T., Ardiç E., Sayilirbaş, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen–host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* 43, D583–D587. doi: 10.1093/nar/gku1121

Guthke, R., Möller, U., Hoffmann, M., Thies, F., and Töpfer, S. (2005). Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics* 21, 1626–1634. doi: 10.1093/bioinformatics/bti226

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinform.* 11(Suppl. 6): S16. doi: 10.1186/1471-2105-11-S6-S16

Rogers, L. D., Brown, N. F., Fang, Y., Pelech, S., and Foster, L. J. (2011). Phosphoproteomic analysis of Salmonella-infected cells identifies key kinase regulators and SopB-dependent host phosphorylation events. *Sci. Signal.* 4, rs9. doi: 10.1126/scisignal.2001668

Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., et al. (2012). An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front. Microbiol.* 3:85. doi: 10.3389/fmicb.2012.00085

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 09 April 2015
doi: 10.3389/fmicb.2015.00235

A review on computational systems biology of pathogen–host interactions

Saliha Durmuş[1], Tunahan Çakır[1], Arzucan Özgür[2] and Reinhard Guthke[3]*

[1] *Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey,* [2] *Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey,* [3] *Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute, Jena, Germany*

Pathogens manipulate the cellular mechanisms of host organisms via pathogen–host interactions (PHIs) in order to take advantage of the capabilities of host cells, leading to infections. The crucial role of these interspecies molecular interactions in initiating and sustaining infections necessitates a thorough understanding of the corresponding mechanisms. Unlike the traditional approach of considering the host or pathogen separately, a systems-level approach, considering the PHI system as a whole is indispensable to elucidate the mechanisms of infection. Following the technological advances in the post-genomic era, PHI data have been produced in large-scale within the last decade. Systems biology-based methods for the inference and analysis of PHI regulatory, metabolic, and protein–protein networks to shed light on infection mechanisms are gaining increasing demand thanks to the availability of omics data. The knowledge derived from the PHIs may largely contribute to the identification of new and more efficient therapeutics to prevent or cure infections. There are recent efforts for the detailed documentation of these experimentally verified PHI data through Web-based databases. Despite these advances in data archiving, there are still large amounts of PHI data in the biomedical literature yet to be discovered, and novel text mining methods are in development to unearth such hidden data. Here, we review a collection of recent studies on computational systems biology of PHIs with a special focus on the methods for the inference and analysis of PHI networks, covering also the Web-based databases and text-mining efforts to unravel the data hidden in the literature.

Keywords: pathogen–host interaction, computational systems biology, bioinformatics, omics data, protein–protein interaction, metabolic interaction, gene regulatory network, drug target

OPEN ACCESS

Edited by:
Anna Norrby-Teglund, Karolinska Institutet, Sweden

Reviewed by:
Marcio Luis Acencio, São Paulo State University, Brazil
Peter Schaap, Wageningen University, Netherlands

****Correspondence:***
Saliha Durmuş, Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University, 41400 Gebze-Kocaeli, Turkey
salihadurmus@gtu.edu.tr

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *01 December 2014*
Accepted: *10 March 2015*
Published: *09 April 2015*

Citation:
Durmuş S, Çakır T, Özgür A and Guthke R (2015) A review on computational systems biology of pathogen–host interactions. Front. Microbiol. 6:235. doi: 10.3389/fmicb.2015.00235

Introduction

Infectious diseases are one of the preliminary causes of death worldwide each year. Emerging and reemerging diseases and drug resistant pathogens have made the problem more serious for human beings. Therefore, novel therapeutic strategies, called theranostics, are increasingly investigated to fight the biological threats. These strategic solutions require a systems biological approach with a thorough understanding of the underlying mechanisms of infections by focusing on molecular interactions between pathogenic and host organisms (Morens et al., 2004;

Murali et al., 2011; Guthke et al., 2012; Durmuş Tekir and Ülgen, 2013). Systems biology is an interdisciplinary research field in life sciences focusing on the study of non-linear interactions among biology entities through the integration and combination of biomolecular and medical sciences with mathematical, computational, and engineering disciplines (Kitano, 2002). By modeling biological phenomena, systems biology uses a more holistic approach based on omics data instead of the traditional reductionism focusing at only a few molecules and interactions. The pathogen–host interactions (PHIs) may be between proteins, nucleotide sequences, metabolites, and small ligands. The protein–protein interactions (PPIs) have been identified as the most important type in the functioning of PHI systems and therefore are the most studied type (Stebbins, 2005; Korkin et al., 2011; Zoraghi and Reiner, 2013). However, non-coding RNAs (ncRNAs) and metabolites have also been reported to have critical functional roles in virus–host and bacteria–host interactions, respectively (Gottwein and Cullen, 2008; Skalsky and Cullen, 2010; Eisenreich et al., 2013; Saayman et al., 2014).

Different levels of omics data collected from pathogens and/or infected cells are crucial components that drive bioinformatic analyses facilitating the construction and analysis of infection-specific gene-regulatory, metabolic, and protein–protein networks (Westermann et al., 2012; Schulze et al., 2015). Such network-based computational systems biology analyses of PHI-based omics data enable the elucidation of infection mechanisms and their dynamics, the identification of potential drug targets for the next-generation antimicrobial therapeutics, and the development of novel and personalized strategies for the prevention and treatment of infections. With an increasing amount of experimental PHI data, Web-based databases were developed to derive and provide pathogen–host interactome data, usually focusing on specific pathogens or hosts (Wattam et al., 2014; Ako-Adjei et al., 2015; Calderone et al., 2015; Guirimand et al., 2015). Although the available databases are promising in data archiving, a huge amount of PHI data is not stored in any of these databases, since these data are buried in the literature. Therefore, there is an urgent need for novel text mining methods specific for PHI data retrieval. In this paper, the efforts on the collection of PHI-based omics data are reviewed first. Next, a review of the computational systems biology analyses of three major types of PHI networks is provided. Then, the available PHI databases and the current snapshot of the literature on text mining for PHI data are presented.

Omics Data Reflecting PHI Networks

The systems biology approaches with genome-wide molecular profiling using high-throughput techniques to generate omics data are changing the face of infection biology together with the computational methods for heterogeneous data management and integrative analysis via mathematical modeling (Guthke et al., 2012; Law et al., 2013). New insights in the microbial and viral pathogenesis, in particular in the host's immune response to contact with pathogens, offer opportunities for better diagnostics, therapeutics, and vaccines. Thus, systems biology of infection allows to yield novel therapeutic targets (Sarker et al., 2013) and to establish individualized or personalized medicine. The integrative personal omics profile (iPOP) combines genomics, transcriptomics, proteomics, metabolomics, and autoantibody profiles from a single individual over a 14-month period (Chen et al., 2012; Li-Pook-Than and Snyder, 2013).

There are various platforms for handling of measured data from samples, data storage and exchange, data pre-processing and data analysis. Powerful platforms for data management in systems biology have recently become available and are standardized step by step by the Functional Genomics Data Society[1] (FGED, founded in 1999 as MGED; Brazma et al., 2006). Several systems biology projects in Europe including the ones dedicated to PHI research use the SysMO-DB/SEEK system for sharing data, knowledge (including Standard Operating Procedures – SOPs) and mathematical models[2] (Wolstencroft et al., 2011). For the management of genomics, transcriptomics, and (2D-gel) proteomics data in infection research, the data warehouse 'OmniFung' was established to support research on fungi–host interactions[3] (Albrecht et al., 2011, 2007).

The free, open source and open development software project Bioconductor, which is primarily based on the statistical R programming language, provides 934 software packages, 894 annotation and 224 experimental data sets for the bioinformatic analysis and comprehension of high-throughput genomic data[4] (Version 3.0). These packages as well as other R packages not included in the Bioconductor project are useful for the advanced, in particular integrative, analysis of omics data and modeling of PHIs. To identify genes, proteins or metabolites of interest for biomarker discovery or drug target prediction by supervised machine learning methods, there are many data mining tools available. For instance, WEKA[5] or RapidMiner[6] is used to characterize the response of the host immune system by decision tree analysis of flow cytometric data (Simon et al., 2012). In addition, there are platforms and software tools for the integrative and explorative analysis and visualization of data from the different omics levels of PHIs (Horn et al., 2014).

PHI-Based Genome and Transcriptome Data

The genomic information from the host and the pathogen represents the basis for all further molecular analyses and bioinformatic investigations of PHI systems. Thus, genome sequencing is fundamental. It helps to improve diagnosis, typing of pathogen, virulence and antibiotic resistance detection, and development of new vaccines and culture media. Single nucleotide polymorphism (SNP) typing is important for both identification and characterization of variants of pathogens (strains, clinical isolates) as well as to study the susceptibility of humans for certain infections. In the last decade, there was, and in the future there will be, an explosion of genome sequence data.

[1] http://fged.org
[2] www.sysmo-db.org
[3] www.omnifung.hki-jena.de
[4] http://bioconductor.org
[5] http://www.cs.waikato.ac.nz/ml/weka
[6] www.rapidminer.de

The new sequencing technologies enable small research units to create huge genome datasets at low cost in short time. As a result, handling, comparing, and extracting useful information from millions of sequences becomes more and more challenging, i.e., increased efforts in computational biology are urgently needed. In particular, sequencing is used for genomic and transcriptomic characterization of new emerging pathogens. Whole-genome sequencing based phylogenetic studies have implications for understanding the evolution of the PHIs as well as tracking and possibly preventing infection diseases as performed for the Enterotoxigenic *Escherichia coli* (ETEC), a major cause of infectious diarrhea (von Mentzer et al., 2014). Metagenomic and metatranscriptomic studies of pathogens revealed how pathogenic microorganisms adapt to hosts, e.g., plants (Guttman et al., 2014).

The first step of genome sequence analysis, the assembling of genome sequence data into a single genomic contig, may be difficult, in particular due to assembling repeated sequences if reference genomes are not available. Then, additional information may be required to resolve the remaining DNA regions. The next step, the functional annotation of virulence-relevant pathogens and focusing on host-interaction genes, is often difficult as the genes of interest for PHIs are frequently species-specific and, thus, studies of gene homologies may not be helpful. The situation would be improved by the databases of protein families involved in host interactions, which incorporate the currently used gene names, sequence motifs, gene functions, and experimental results (see section "Web-Based Databases for PHI Systems"). On the other hand, comparative genomics can provide insights into molecular pathogenesis, host specificity, and evolution of pathogens. Next generation sequencing (NGS) has revolutionized the molecular investigation of the diversity of pathogens on the genomic and transcriptomic level. It enables an efficient analysis of complex human microfloras, both commensal and pathological, through metagenomic methods. Genomic sequences and their annotations are provided through several portals, such as the Genomes Online Database[7].

In contrast to the static information from the genome, the transcriptome reflects the dynamics of PHI systems that results in temporal profiles of gene expression with changes in the scale of minutes and hours. More and more, beside the protein-coding mRNAs, also various non-conding small RNAs are investigated. For instance, in *Staphylococcus aureus*, a leading pathogen for animals and humans, about 250 regulatory RNAs were found (Guillet et al., 2013). Repositories for transcriptome data, such as Gene Expression Omnibus[8] (GEO) and ArrayExpress[9] freely distribute microarray and NGS (RNA-Seq) data as well as other forms of high-throughput functional genomics data. In GEO, data from more than 1600 organisms, both pathogens and hosts, are accessible. For instance, for the pathogens *Mycobacterium tuberculosis*, *S. aureus*, *Candida albicans*, and *Helicobacter pylori* transcriptome data from 1,855, 1,777, 1,627, and 1,284 samples are available, respectively. Other data sets monitor the transcriptome of the host's response, e.g., *Homo sapiens* and *Mus musculus* (GSE56091, GSE56093). Some monitor data from host and pathogen simultaneously, e.g., *S. aureus* and the zebrafish *Danio rerio* (GSE32119). NGS has opened the door for simultaneous transcriptome analysis by the so-called dual RNA-Seq (Tierney et al., 2012a,b; Westermann et al., 2012; Camilios-Neto et al., 2014; Longo et al., 2014; Pittman et al., 2014; Xu et al., 2014; Schulze et al., 2015).

PHI-Based Proteome and Metabolome Data

Proteins are key players in PHIs, in particular in pathogen recognition as well as innate and adaptive immune responses. Pathogen-associated molecular patterns (PAMPs) are molecules or small molecular motifs within a group of pathogens (e.g., the protein flagellin, lipopeptides, lipopolysaccharide – LPS) that are recognized by proteins, the so-called pattern recognition receptors (PRRs), such as Toll-like receptors (TLRs; Qian and Cao, 2013). For instance, TLR4 recognizes bacterial LPS, and TLR5 recognizes bacterial flagellin. The PRRs stimulate signal transduction via pathways, e.g., the tumor necrosis factor alpha (TNFα) signaling or the interferon-gamma (IFNγ)-receptor pathway including the JAK-STAT-pathway. IFNγ is a cytokine that is a key player in innate and adaptive immunity against viral, as well as some microbial and protozoan infections. The nuclear factor NF-κB is a protein, a transcription factor, that is activated by various intra- and extra-cellular stimuli such as bacterial or viral products, for instance via the TLRs signaling and induces the expression of pro-inflammatory cytokines (interleukines, TNFα, Type I interferones). Thus, the application of proteomics is crucial in the investigation of PHI systems and for the above mentioned iPOP, e.g., the immune profiling of patients (Chen et al., 2012).

By dedicated bioinformatic pipelines, a description of pathogen proteomes and their interactions within the context of human host has a strong impact in both diagnostic and clinical treatment of the patient. In the last few years, several advanced proteomic techniques have been established providing individual proteome charts of both pathogens and hosts, including antimicrobial or antimycotic resistance profiling and immune profiling of the patient. Proteome analysis is hampered by the extremely divergent biochemical properties of the individual proteins, making an entire view of the proteome almost impossible (Otto et al., 2014). The coupling of multidimensional separations with mass spectrometry (MS) for protein and peptide analyses via, for instance, the matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) techniques resulted in powerful MS instrumentations. Many of these MS-based techniques, e.g., MALDI-TOF, have been used in clinical microbiology and research (Del Chierico et al., 2014; Otto et al., 2014). For PHI analyses, the cell wall proteins and the secretomes are of special interest to study the PAMPs and PRRs as well as their interplay (Schmidt and Völker, 2011; Zheng et al., 2011; Heilmann et al., 2012; Di Carli et al., 2012). PHI analysis studies that focus on the host side studying the immune response (Hartlova et al., 2011; Heyl et al., 2014) or on the pathogen side

[7]https://gold.jgi-psf.org

[8]http://www.ncbi.nlm.nih.gov/geo

[9]https://www.ebi.ac.uk/arrayexpress

(Bröker and van Belkum, 2011; Cash, 2011; Ahmad et al., 2012) have also been conducted. The integrative analysis of proteome data with other omics data for both pathogens and hosts is a very challenging task in bioinformatics (Albrecht et al., 2010, 2011).

Stanberry et al. (2013) demonstrated on the host side a strong association between the metabolome profiles, i.e., the metabolite expression levels of differentially expressed pathways, and their temporal patterns at each time point with the disease status of viral infection with a human rhinovirus and a respiratory syncytial virus. For metabolic studies on the pathogen side, there are *in silico* strategies to identify effective targets for anti-infective drugs based on constraint-based modeling of genome-scale metabolic networks (Chavali et al., 2012; see section "PHI Metabolic Network Models"). A prominent type of PHIs is the production of toxins by the pathogens that attack the host. For instance, gliotoxin produced by the human-pathogen fungus *Aspergillus fumigatus* modulates the immune response and induces apoptosis in the host (Gardiner and Howlett, 2005; Scharf et al., 2012). Another type of PHI is due to the pathogens that frequently utilize substrates from the host (Rohmer et al., 2011). The gene regulatory network (GRN) model-assisted studies of the uptake of essential substrates such as iron (Linde et al., 2010, 2012) or nitrogen sources (Ramachandra et al., 2014) by such pathogens address specific but important aspects of PHIs.

Computational Systems Biology of PHI Networks

A systems biology approach is crucial to model and understand PHIs, in particular interactions between the immune system of humans or animals, and the pathogens (Berglund et al., 2009; Guthke et al., 2012; Horn et al., 2012; Zhou et al., 2013). Systems biology of PHIs aims at describing and analyzing the confrontation of the host with viral, bacterial, and fungal pathogens and parasites by the development of testable computational models of PHIs. The predictive power of such models enables diagnosis and therapy by the prediction of biomarkers and drug targets. Systems biology of PHIs includes an integrative analysis and modeling of genome-wide and/or spatio-temporal data from both the host and the pathogen, or the response of the host or pathogenic cells to defined perturbations that simulate conditions during infection.

At the computational side, systems biology of PHIs comprises:

- Modeling of molecular mechanisms of infections,
- Modeling of non-protective and protective immune defenses against pathogens to generate information for possible immune therapy approaches,
- Modeling of PHI dynamics and identification of biomarkers for diagnosis and for individualized therapy of infections,
- Identifying essential virulence determinants and host factors, and thereby predicting potential drug targets
- Understanding of PHIs, in particular the immune system and the immune evasion of the pathogens, as the result of evolutionary long-term adaptation and selection.

Both the innate and the adaptive immune system comprise cell-mediated and humoral components. Thus, systems biology of immune defense has to handle multi-scale modeling from molecular to systemic/organ level. The same is required for the pathogen side. The interaction of cellular components is preferentially the area of the agent-based modeling, whereas the humoral immunity can be modeled by ordinary differential equations (ODEs). While the innate immune response is non-specific and acts immediately, the adaptive immune response is pathogen and antigen specific with time lag and immunological memory. Thus, the temporal organization and population dynamics have to be modeled in a different manner for the innate and adaptive immune system in interaction with the pathogen (Perelson, 2002; Gottschalk et al., 2013; Six et al., 2013; Panayidou et al., 2014).

The study of the interplay between pathogens and immune cells remains a challenging task due to its complexity. While the emerging image-systems biology of cellular interaction (Mech et al., 2011; Hünniger et al., 2014; Kraibooj et al., 2014; Pollmächer and Figge, 2014) is here out of the scope, the present review focuses on the molecular, mainly omics data-based level. Here, a difficulty arises to separate host's transcripts, proteins, and metabolites from that in the pathogen and to extract them in a balanced amount for a simultaneous monitoring of these molecules so that the network models of PHIs are inferred. Therefore, most studies focus either on the pathogen or the host side with a defined and controlled change of the respective other side as an external perturbation, i.e., considering an input from the outside of the investigated system. Thus, to simplify the study, the PHIs have been studied mainly in one direction either from pathogen to host or from host to pathogen. Only very recently, the bi-directional interaction of pathogen and host became observable simultaneously using the so-called dual RNA-Seq data generated by NGS of the transcriptome of pathogen and host (see section "PHI-Based Genome and Transcriptome Data").

Understanding the evolutionary dynamics of PHIs by mathematical modeling in terms of both molecular mechanisms and selective forces is important in order to design drugs that will be effective in the long term, i.e., to avoid or to overcome resistance to antibiotics (Guo et al., 2011; Lima et al., 2013; Palmer and Kishony, 2013). Finally, computational systems biology approaches are and will be used to select pathogen-host drug targets and to develop novel anti-infectives and vaccines (Brown et al., 2011; Mooney et al., 2013; Sarker et al., 2013; Rienksma et al., 2014).

PHI Regulatory Network Models

Biological network models are widely used to improve our understanding of infectious diseases (Mulder et al., 2014). There are many small-scale models (mainly ODE-based), which describe PHIs phenomenologically (Baccam et al., 2009; Saenz et al., 2010; Manchanda et al., 2014). These models without molecular specification are out of the scope of this review, as they usually do not predict PHIs on the molecular level. Here, omics data based PHI models will be reviewed.

Computational modeling of GRNs reveals the molecular logic of adaptation of pathogens to their hosts, the immune evasion of

the pathogen as well as the immune response of the host to infection with pathogens. GRNs provide causal explanations for the differentiation, the developmental and effector states, as well as the fate dynamics of immune cells (Singh et al., 2014). Finally, GRNs may also describe the interaction of the two networks, one of the pathogen and the other of the host (see **Figure 1** for example). The inference of GRN models from gene expression data is a problem of great importance for PHI studies. Various reverse engineering methods have been proposed, which include methods based on Boolean networks, Bayesian networks, differential or difference equations, and graphical Gaussian models. In general, due to the high dimensionality (thousands of genes and proteins in both host and pathogen organisms) versus the limited number of samples (not more than hundreds in the case of steady state data from knock-out (KO) mutants; only a few samples in *in vivo* studies of PHI monitored at, e.g., 5–10 time points), the GRN inference is underdetermined implying that there could be many equivalent (indistinguishable) solutions. Motivated by this fundamental limitation, there are various approaches for GRN inference. Again, there are outstanding review articles covering the long-standing problem of gene expression data-driven GRN inference (De Jong, 2002; van Someren et al., 2002; Gardner and Faith, 2005; Bansal et al., 2007; Emmert-Streib et al., 2014; Linde et al., 2015). One of the conclusions from the DREAM initiative[10] (Dialog for Reverse Engineering Assessment of Methods; Prill et al., 2010) that performed a comprehensive blind assessment of over 30 network inference methods was that no single inference method performs optimally across all datasets. Integration of predictions from multiple inference methods shows more robustness and higher performance across diverse datasets (Marbach et al., 2012). For instance, the algorithm TRaCE performs an ensemble inference of GRNs, which takes into account the inherent uncertainty associated with discriminating direct and indirect gene regulations from steady-state data of KO experiments (Ud-Dean and Gunawan, 2014). Another group of GRN inference approaches includes prior knowledge as reviewed by (Hecker et al., 2009; Isci et al., 2014) or further experimental data (Greenfield et al., 2010). A third group of GRN algorithms restricts the GRN to static networks inferred from steady state data (e.g., from KO mutants of the pathogen) or to small-scale networks with a few nodes (genes, proteins), where the pre-selection of them is the critical point (Nakajima and Akutsu, 2014).

The genome-wide GRN model inference, when restricted to the static network models of thousands of genes, requires large gene expression data sets and prior-knowledge in high quality and quantity, which is not the case for most of the pathogens of interest as demonstrated for the human-pathogen *C. albicans* (Altwasser et al., 2012). In contrast to the genome-wide GRN models, the small-scale network models that take into account 5–50 genes or proteins are often used for PHI studies. These models do not represent the holistic view as it is claimed in systems biology, but they generate hypotheses of PHIs that drive further experimental work in infection biology. Afterward, the GRN-based *in silico* predictions have to be validated experimentally. This approach of focused small-scale GRN

[10] www.the-dream-project.org

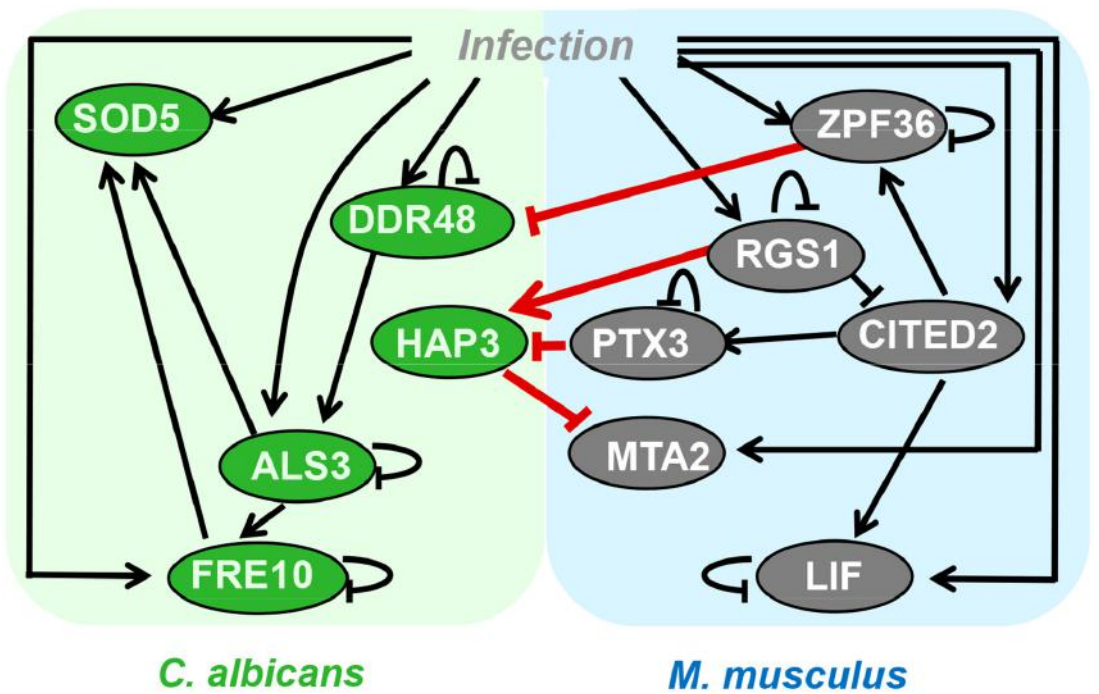


FIGURE 1 | Network model describing pathogen-host interactions between *C. albicans* and murine dendritic cells based on dual RNA-Seq data (modified from Tierney et al., 2012b).

inference was reported particularly for human–pathogen fungal infection (Linde et al., 2010, 2012; Ramachandra et al., 2014) by using the ODE-based NetGenerator algorithm. The algorithm was primarily introduced to model the immune response to bacterial infection (Guthke et al., 2005; Weber et al., 2013). This algorithm was also applied for the inference of the PHIs of the human–pathogen fungus *C. albicans* with murine dendritic cells based on dual RNA-Seq data (Tierney et al., 2012b). Here for instance, based on the inferred GRN model shown in **Figure 1**, an inhibition of the expression of the protein HAP3 in the fungus by the murine pentraxin (PTX3) was computationally predicted and, afterward, experimentally validated.

PHI Metabolic Network Models

Pathogens are dependent on the host environment for the substrates required to maintain a metabolically active state (Chavali et al., 2012; Eisenreich et al., 2013). Therefore, the exchange of several metabolites takes place between pathogens and their host. Besides, the production of virulence factors by the pathogen requires energy, and, hence, an active metabolism, making the nutrients in the host environment crucial for the infection to occur (Milenbachs et al., 1997). The direct functional link between metabolism and virulence is also supported by the finding that metabolic and virulence genes are located on the same pathogenicity island for some pathogens (Rohmer et al., 2011; Heroven and Dersch, 2014). In a different approach, the authors used a network-based computational analysis to elucidate common targeting strategies of bacteria and viruses on human (Durmuş Tekir et al., 2012), based on pathogen–host PPIs stored in the PHISTO database (Durmuş Tekir et al., 2013). Their results revealed metabolism as a common strategy of both pathogen types to target human cells. The role of metabolism in the pathogenesis was also emphasized by others (Kafsack and Llinás, 2010). Therefore, metabolism is a candidate target for anti-microbial therapies.

There are well-established bioinformatic methods for metabolic network reconstruction, based on DNA genome sequences and constraint based modeling covered by outstanding review articles (Feist et al., 2008; Oberhardt et al., 2009; Ruppin et al., 2010; Bordbar and Palsson, 2012). The *in silico* methods for metabolic network reconstruction are highly valuable for understanding the physiology of the pathogen, e.g., the biosynthesis of toxins that attack the host or the substrate requirement that shows the dependency of the pathogen on the environment within the host. At the host side, the human metabolic network reconstruction may also have an impact for drug discovery and development (Ma and Goryanin, 2008). A systematic modeling of the metabolic trafficking between pathogens and its hosts first started with the constraint-based modeling of the Gram-negative bacterial pathogen, *Salmonella typhimurium* (Raghunathan et al., 2009). The authors reconstructed a genome-scale metabolic model for the pathogen in question, and then simulated its survival capabilities with the flux-balance approach (Kauffman et al., 2003; Orth et al., 2010). When they used a media mimicking host-cell nutrient environment (e.g., macrophage) rather than laboratory media, their correct predictions considerably increased. They also showed that the use of gene expression data can lead to a better inference of active transport mechanisms, and hence the host cell environment. In another study, the reconstructed metabolic network of the malaria-causing protozoan parasite, *Plasmodium falciparum*, was embedded into its host, erythrocyte, and the combined pathogen-host network was simulated via flux-balance analysis (FBA; Huthmacher et al., 2010). The novelty here was to take also the host network into account to predict metabolite exchanges between the parasite and the host, rather than only considering the host environment to account for pathogen–host metabolic interactions. Such a consideration is important since a pathogen infection causes pathogen-specific or common responses in the host metabolic pathways from central carbon metabolism to fatty acid and amino acid metabolisms (Eisenreich et al., 2013). Their analysis resulted in the prediction of antimalarial drug targets (Huthmacher et al., 2010).

In a more systematic study, genome scale metabolic networks of *Mycobacterium tuberculosis* and its host, alveolar macrophage, were reconstructed in an integrated fashion and the integrated pathogen-host metabolic model was used to analyze infection mechanisms and related different pathological states (Bordbar et al., 2010). The reconstructed joint metabolic network covered 2071 genes (661 for the pathogen, 1410 for the macrophage), controlling a total of 4489 reactions. Integrative analysis of the network with the transcriptome data from the infected macrophage cells enabled the inference of the induced changes in the pathogen. One important issue in the network based drug-target identification is the selectivity of the identified targets. The candidate target must make no harm to the host. This was taken into consideration by (Bazzani et al., 2012), where they used the integrated pathogen-host metabolic model of *Plasmodium falciparum* and hepatocyte, the first human infection site for malaria parasites. The flux balance approach was combined with 48 experimental antimalarial drug targets to identify the targets which are essential for the parasite but not essential for hepatocyte metabolism. The *in silico* analysis led to the ranking of the identified targets with respect to their reducing effect on the cellular fitness.

One key point in the elucidation of metabolic mechanisms both in the host and in the pathogen is to correctly characterize the nutrient availability for the pathogen in the host environment. This characterization is also important for successful modeling attempts. The available nutrients shape the active parts of the pathogen metabolism, and also the depletion of different metabolites may trigger different responses in the host (Bumann, 2009; Rohmer et al., 2011; Eisenreich et al., 2013; Sasikaran et al., 2014). Therefore, nutritional environment has a crucial role to understand the basis of infection mechanisms (Brown et al., 2008; Gouzy et al., 2014). Systems-level experimental approaches such as lipidomics and metabolomics are getting popular to decipher the pathogen–host nutritional interactions (Wenk, 2006; Olszewski et al., 2009; Antunes et al., 2011). A recent attempt to identify active metabolic routes from the host environment to pathogen inside by using 13C flux spectral analysis (Beste et al., 2013) provided a quantitative measure of interactions between

Mycobacterium tuberculosis and its host macrophage. The experimental labeling data enabled the identification of substrates used by the pathogen. Another elegant study used 13C-labeling based fluxomics as well as metabolomics and proteomics to shed light on the metabolic interplay between *Shigella flexneri* and HeLa epithelial cells (Kentner et al., 2014). They were able to identify host metabolites that contribute to the growth of Shigella as substrates.

Similar to the use of gene expression data to infer GRNs as discussed in the previous section, metabolome data obtained from the infected cells or PHI systems can be used to infer infection-specific metabolic networks by using reverse engineering approaches. Taking into account several bioinformatics methods proposed for this type of inference as reviewed recently (Cakir and Khatibipour, 2014), we believe the field of infection will witness promising applications in the coming years.

PHI Protein–Protein Network Models

In the post-genomic era, genes and the corresponding proteins are studied thoroughly, allowing the identification of intra- and interspecies protein interaction networks. Following the development of experimental techniques to produce large-scale molecular interaction data (Fields and Song, 1989; Fisher et al., 2002; Gavin et al., 2002; Ho et al., 2002), the first large-scale intraspecies PPIs were produced experimentally (Finley and Brent, 1994; Bartel et al., 1996; Fromont-Racine et al., 1997; Flajolet et al., 2000; Ito et al., 2000; McCraith et al., 2000; Walhout et al., 2000; Rain et al., 2001). On the other hand, the initial efforts to identify large scale interspecies protein interaction data for PHI systems have been performed since 2007 (**Table 1**). The first large scale PHI examples were for commonly observed and human-threatening viruses and bacteria. These were firstly for viral pathogens; Epstein-Barr virus (EBV; Calderwood et al., 2007; Forsman et al., 2008), Hepatitis C virus (HCV; De Chassey et al., 2008; Tripathi et al., 2010; Dolan et al., 2013; Ngo et al., 2013), Human Immunodeficiency Virus (HIV; Gautier et al., 2009; Jäger et al., 2012), Influenza A virus (Shapira et al., 2009), Dengue virus (DENV; Khadka et al., 2011), Measles virus (MV; Komarova et al., 2011), and Human Respiratory Syncytial Virus (HRSV; Wu et al., 2012). On the other hand, the large scale experimental detection of bacteria-human protein interaction networks was performed for *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis* (Dyer et al., 2010; Yang et al., 2011).

As an initial large scale virus–human PHI network example, protein interactions between the herpesvirus EBV and human were mapped by the yeast two hybrid (Y2H) method, providing 173 PHIs between 40 EBV proteins and 112 human proteins (Calderwood et al., 2007). EBV is the infectious cause of several human diseases such as Burkitt's lymphoma, Hodgkin's disease, and nasopharyngeal carcinoma. This EBV–human protein interaction network enabled the initial observations about EBV strategies (i.e., targeting hub and bottleneck human proteins) for replication and persistence within the host. For the same viral system, 147 human protein interactors for EBV nuclear antigen 5 (EBNA5) were identified with LC-MS/MS in a following study (Forsman et al., 2008). Multifunctional viral protein EBNA5 is already known to be critical in EBV pathogenesis, and these PHI data provided further insights on its molecular mechanisms during infection. The identified interactions between EBNA5 and the human proteins functioning in protein control systems that recognize proteins with abnormal

TABLE 1 | The large-scale pathogen–human PPI networks in chronological order.

Pathogen name	Pathogen type	Number of PHIs	Number of interacting pathogen proteins	Number of interacting human proteins	Reference
EBV	DNA virus	173	40	112	Calderwood et al. (2007)
HCV	RNA virus	481	11	421	De Chassey et al. (2008)
EBV	DNA virus	147	1	147	Forsman et al. (2008)
HIV-1	Retrovirus	183	1	183	Gautier et al. (2009)
Influenza A virus (H1N1 A/PR/8/34)	RNA virus	135	10	87	Shapira et al. (2009)
Influenza A virus (H3N2 A/Udorn/72)	RNA virus	81	10	66	Shapira et al. (2009)
Bacillus anthracis	Gram-positive bacteria	3,073	943	1,748	Dyer et al. (2010)
Yersinia pestis	Gram-positive bacteria	4,059	1,218	2,108	Dyer et al. (2010)
Francisella Tularensis	Gram-negative bacteria	1,383	349	999	Dyer et al. (2010)
HCV	RNA virus	56	2	56	Tripathi et al. (2010)
DENV	RNA virus	139	10	105	Khadka et al. (2011)
MV	RNA virus	245	1	245	Komarova et al. (2011)
Y. pestis	Gram-positive bacteria	204	66	109	Yang et al. (2011)
HIV-1	Retrovirus	497	16	435	Jäger et al. (2012)
30 viral species	DNA and RNA viruses	1681	70	579	Pichlmair et al. (2012)
HRSV	RNA virus	221	1	221	Wu et al. (2012)
HCV	RNA virus	112	7	94	Dolan et al. (2013)
HCV	RNA virus	103	1	103	Ngo et al. (2013)

structures may indicate the roles of the viral protein in this system.

The first proteome-wide PHI map for the flavivirus HCV, a major cause of chronic liver diseases, was deduced by Y2H and then by literature mining of previously found interactions between HCV and human, providing a large network for such a small-genome organism. The resulting network consists of 481 interactions between 11 HCV proteins and 421 human proteins. Pathway enrichment analysis of the targeted cellular proteins indicated focal adhesion as a new function subverted by HCV (De Chassey et al., 2008). Using the same experimental approach, 11 human proteins interacting with HCV Core protein and 45 interacting with NS4B (one of the six HCV non-structural proteins) were found (Tripathi et al., 2010). To further understand the mechanisms of the interactions between HCV and human proteins, two extended PPI networks were constructed. These networks are composed of the Y2H-derived interactions and the secondary interactors of the human proteins that interact with the Core and NS4B proteins. Functional analysis of these networks pointed to the human proteins ENO1, SLC25A5, and PXN as potential antiviral targets. ENO1 and SLC25A5 are interaction partners of HCV Core protein. PXN is the first neighbor of both ENO1 and SLC25A5 within the human PPI network. Observing the effects of small interfering RNA (siRNA) knockdown of these host proteins on HCV propagation and replication validated the computational network analysis results (Tripathi et al., 2010). Another Y2H screen resulted in 112 unique interactions between 7 HCV and 94 human proteins (Dolan et al., 2013). DENV is another member of the flaviviruses family, causing the severe human disease dengue hemorrhagic fever. Using the Y2H method, 139 PHIs were detected between 10 DENV proteins and 105 human proteins (Khadka et al., 2011). These two PHI networks of HCV–human by Dolan et al. (2013) and DENV–human by Khadka et al. (2011) were analyzed comparatively and a large overlap was observed between HCV and DENV targets. To determine if the common cellular targets play crucial roles in infections, siRNA experiments were performed and the results revealed the required cellular proteins (CUL7, PCM1, RILPL2, RNASET2, and TCF7L2) for HCV replication (Dolan et al., 2013). Finally, using protein microarray assays, 103 human proteins were identified as HCV Core-interacting partners. Through these PHI data, the viral modulation of some cellular mechanisms was studied in detail and the cellular MAPKAPK3 was proposed as a potential therapeutic target for HCV infections (Ngo et al., 2013). Prior to these studies, a number of small scale PHI data were produced for the HCV–human interaction system (Matsumoto et al., 1997; Hsieh et al., 1998; Lu et al., 1999; Owsianka and Patel, 1999).

Orthomyxovirus Influenza A virus is the source of all flu pandemics infecting multiple species. For H1N1 A/PR/8/34 strain of influenza virus, 135 PHIs were identified between 10 viral and 87 human proteins, most of which are expressed in primary human bronchial cells. For another strain of influenza A virus, H3N2 A/Udorn/72, a PHI network with 81 interactions between 10 viral and 66 human proteins was constructed. Both of the PHI networks were detected by the Y2H method. Similarities of these two PHI networks highlighted the conserved functions of influenza virus proteins through strains. Observing the topological network properties of these Influenza A virus–human PPI networks allowed to draw crucial conclusions on the multifunctionality of the small number of proteins encoded by RNA viruses, revealing that viral proteins can interact with a significant number of human proteins (Shapira et al., 2009).

AIDS-causing retrovirus HIV, probably the most studied human pathogen, depends largely on human cellular machinery to be replicated, like other RNA-carrying viruses. One large-scale PHI dataset for HIV-1 was produced using affinity chromatography coupled with MS, resulting in 183 human nuclear proteins as interacting partners of HIV-1 Tat (nuclear regulatory protein) which is essential for viral replication within the host nucleus. The following *in silico* analysis of the experimentally verified PHI data provided further insights on the mechanisms of Tat during HIV-1 infection. Firstly, motif composition analysis highlighted that Tat-targeted cellular proteins are enriched for domains mediating protein, RNA and DNA interactions, and helicase and ATPase activities. Secondly, functional analysis of Tat-targeted human proteins showed that they are enriched for a wide range of biological processes such as gene expression regulation, RNA biogenesis, chromatin structure, chromosome organization, DNA replication, and nuclear architecture (Gautier et al., 2009). Another large PHI network was constructed for HIV–human protein complexes by affinity tagging and purification MS, resulting in 497 PHIs between 16 HIV-1 proteins and 435 human proteins. In that study, the functional categories of HIV-targeted human proteins were analyzed indicating that the host factors in the found PHI network are enriched for the transcription and the regulation of ubiquitination. Additionally, the domains of the interacting proteins were also investigated, and the enriched domain types (14-3-3 domains and β-propellers) in targeted human proteins were identified to facilitate future structural modeling studies (Jäger et al., 2012). For HIV-1, several small scale experiments were also carried out to find protein PHI data (Cujec et al., 1997; Le Rouzic et al., 2002; BonHomme et al., 2003; Lusic et al., 2003; Naji et al., 2012) establishing HIV-1 as the pathogenic species having the largest experimentally verified PHI data.

Using the approach of combining modified tandem affinity chromatography and MS analysis, 245 cellular interacting proteins were identified for the viral protein MV-V (one of the virulence factors of paramyxovirus MV). MV-V was found to target known key components of the host antiviral response including STAT1, STAT2, IFIH1, and p53, and also essential components of ribosome, reticulum, and mitochondria. The topological and functional analysis of human proteins targeted by MV-V shows that they have properties within the human interactome similar to the well-known targets of other viruses (Komarova et al., 2011).

As an example for another multi-functional viral protein, HRSV (another member of paramyxoviruses) NS1 can act as an antagonist of host type I and III interferon production and signaling, inhibit apoptosis, suppress dendritic cell maturation, control protein stability, and regulate transcription of host cell mRNAs,

among its other functions. A total of 221 PHIs were determined between only one viral protein NS1 and human proteins, reflecting its multifunctional nature. This virus-human PHI network was produced by quantitative proteomics in combination with green fluorescent protein (GFP)-trap immunoprecipitation. It was observed that many of the HRSV-targeted human proteins have roles in transcriptional regulation and cell cycle regulation (Wu et al., 2012).

A study covering several DNA and RNA viruses (Pichlmair et al., 2012) found 1681 PHIs between 70 viral ORFs from 30 species and 579 human proteins. The interacting cellular proteins were isolated by tandem affinity purification (TAP), and the purified proteins were analyzed by one-dimensional gel-free liquid chromatography tandem MS (LC–MS/MS). A comparative interactomics analysis of the produced viral PHI networks (DNA viruses versus RNA viruses) provided crucial insights on the infection strategies of DNA and RNA viruses. It was concluded that RNA viruses target the JAK–STAT and chemokine signaling pathways, as well as pathways associated with intracellular parasitism, whereas DNA viruses target cancer pathways (Pichlmair et al., 2012).

The first extensive bacterial PHI networks were identified for important human pathogens, *B. anthracis*, *F. tularensis*, and *Y. pestis* (Dyer et al., 2010; Yang et al., 2011). Gram-positive bacteria *B. anthracis* and *Y. pestis* and Gram-negative bacterium *F. tularensis* are respiratory pathogens causing anthrax, bubonic plague, and acute pneumonic disease, respectively. Using the Y2H method, large-scale interaction data were generated between these bacteria and human, leading to 3073 PHIs between 943 *B. anthracis* proteins and 1748 human proteins, 4059 PHIs between 1218 *Y. pestis* proteins and 2108 human proteins, and 1383 PHIs between 349 *F. tularensis* proteins and 999 human proteins. Bioinformatic analysis of these experimentally found bacteria–human interaction data revealed that bacterial proteins preferentially interact with human proteins that are hubs and bottlenecks in the human PPI network, as previously observed for viral PHIs. The modules of bacterial PHIs that are conserved amongst the three networks were computed. The found conserved modules may reveal commonalities among how different bacterial pathogens interact with crucial host pathways involved in inflammation and immunity (Dyer et al., 2010). A different Y2H strategy was used for *Y. pestis* by choosing only potential virulence factors as bait proteins. 204 PHIs were identified between 66 *Y. pestis* proteins and 109 human proteins, and then 23 previously reported PHIs were integrated to construct a comprehensive network between *Y. pestis* and human (Yang et al., 2011).

The increase in the amount of experimentally verified pathogen–human PPI data allowed a number of bioinformatic studies to investigate infection mechanisms at the level of PHIs for different pathogen types (Dyer et al., 2008; Singh et al., 2010; Durmuş Tekir et al., 2012). The first global analysis of more than 10,000 PHI data revealed important observations (Dyer et al., 2008). Firstly, targeting hub and bottleneck proteins were concluded as a common behavior for all pathogens. Targeting human transcription factors and key proteins that control the cell cycle and regulate apoptosis and transport of genetic material across the nuclear membrane were found to be common infection strategies of viruses. On the other hand, targeting human proteins that function in the immune response was observed as a common bacterial infection strategy (Dyer et al., 2008). In a following study, investigation of more than 20,000 experimental PHI data revealed that the preference of interacting with hub and bottleneck proteins is more pronounced in viruses than bacteria. The analysis of the human proteins targeted by both bacteria and viruses indicated that attacking human metabolic processes is a common strategy used by both pathogens (Durmuş Tekir et al., 2012). In addition to these comparative interactomics studies for bacterial and viral PHI networks, a comparative analysis of virus interactions with human signal transduction pathways revealed that different viruses tend to target the same cellular pathways, not necessarily via interacting with the same cellular proteins (Singh et al., 2010).

Web-Based Databases for PHI Systems

In parallel with the first large-scale experimentally verified PHI data, the initial efforts on the development of PHI-specific databases were performed toward the end of the first decade of this century (**Table 2**). Currently, a number of Web-based resources aim to integrate pathogen–host molecular interactions and related data available in the literature. Some of them store data on only one specific pathogen species as in the case of HCVpro (Kwofie et al., 2011), HIV-1 Human Interaction Database at NCBI (Ako-Adjei et al., 2015), HoPaCI-DB (Bleves et al., 2014) for *Pseudomonas aeruginosa* and *Coxiella burnetii*, and Proteopathogen (Vialás et al., 2009) for *C. albicans*. The resources based on a wider range of specific pathogens are VirHostNet (Guirimand et al., 2015), VirusMentha (Calderone et al., 2015) and ViRBase (Li et al., 2015) for viruses, PATRIC (Wattam et al., 2014) for bacteria and PHI-base (Urban et al., 2015) for bacterial, fungal, and oomycete pathogens. Finally, PHIDIAS (Xiang et al., 2007), HPIDB (Kumar and Nanduri, 2010), and PHISTO (Durmuş Tekir et al., 2013) are PHI databases for all pathogen types with known interaction data.

HCVPro (HCV interaction database) is dedicated to only HCV, cataloging the characterized protein interactions for intraviral and virus–human systems. Additionally, it includes information on the structure and functions of HCV proteins (Kwofie et al., 2011). The HIV-1 Human Protein Interaction Database at NCBI includes the interactions between HIV-1 and human proteins. In its content, the majority of the protein interaction data are indirect (e.g., upregulation, modification) whereas the rest are direct (e.g., binding; Ako-Adjei et al., 2015). HoPaCl-DB (Host–Pseudomonas and Coxiella interaction database) provides information on interactions between molecules, bioprocesses, and cellular structures for the bacterial pathogens *Pseudomonas aeruginosa* and *C. burnetti* and their host organisms. The graphical representation of these interaction systems is also available in HoPaCl-DB (Bleves et al., 2014). The other pathogen-specific data resource, Proteopathogen is a protein database for studying *C. albicans*–host interactions.

TABLE 2 | Contents of Web-based PHI databases.

Database	Number of PHIs	Pathogen	Host	Reference
HCVPro	524	Only HCV	Only human	Kwofie et al. (2011)
HIV-1 Human at NCBI	12,786	Only HIV-1	Only human	Ako-Adjei et al. (2015)
HoPaCI-DB	4203	*Pseudomonas aeruginosa* and *Coxiella burnetii*	Mammalia, *Caenorhabditis elegans*, *Drosophila Melanogaster*, *Danio rerio*	Bleves et al. (2014)
HPIDB	40,611	Bacteria, fungi, viruses	Animal, human, plant	Kumar and Nanduri (2010)
PATRIC	8547	Only bacteria	Actinopterygii, Arachnida, Chromadorea, Insecta, Mammalia	Wattam et al. (2014)
PHI-base	4102	Bacteria, fungi, oomycete	Animal, human, insect, fish, fungi, plant	Urban et al. (2015)
PHIDIAS	NA	Bacteria, viruses, parasites	All hosts	Xiang et al. (2007)
PHISTO	39,166	Bacteria, fungi, Protozoa, viruses	Only human	Durmuş Tekir et al. (2013)
Proteopathogen	NA	*Candida albicans*	Mammalia	Vialás et al. (2009)
ViRBase	NA	Only viruses	All hosts	Li et al. (2015)
VirHostNet	16,000	Only viruses	Animal, human, plant	Guirimand et al. (2015)
VirusMentha	8084	Only viruses	All hosts	Calderone et al. (2015)

Although the focus of the database is on *C. albicans* and its interactions with macrophages, the database also includes data for different fungal pathogens and other mammalian cells. Proteopathogen provides additional information about the interacting proteins such as Gene Ontology (GO) and pathway annotations, and protein structures (Vialás et al., 2009).

PATRIC (The PathoSytems Resource Integration Center) is a dedicated resource for bacterial systems including comprehensive data on genomics, transcriptomics, PPIs, 3D protein structures, and sequence typing. However, its focus is on the genomic data, currently covering more than 10,000 bacterial genome sequences. PATRIC provides a private workspace for each user where they can store their own data. In their workspaces, users can perform comparative genomics and transcriptomics via the corresponding analysis tools. PATRIC provides bacteria–host PPI data through its tool Pathogen Integration Gateway (PIG; Wattam et al., 2014). PHI-base (Pathogen–Host Interactions Database) is a Web-accessible PHI database specific for bacterial, fungal, and oomycete pathogens, which are medically and agronomically important. PHI-base serves options to facilitate the discovery of genes that may be potential targets for chemical intervention, containing information on the pathogenicity/virulence genes functioning in the PHI systems. As a genomic data focused resource, PHI-base has the functionalities allowing functional annotations of the genes and comparative genomics analysis (Urban et al., 2015). On the other hand, there are databases developed specifically for viral PHI systems such as VirHostNet (Guirimand et al., 2015), VirusMentha (Calderone et al., 2015) and ViRBase (Li et al., 2015). VirHostNet (Virus–Host Network) is one of the earliest PHI resources specialized in the management and analysis of integrated virus–virus, virus–host, and host–host protein interaction networks coupled to their functional annotations. The host organism in the VirHostNet is only human. Its Web interface provides both table-based and graph-based visualizations of the PHI networks (Guirimand et al., 2015). The recently developed tool, VirusMentha is another virus-virus and virus–host protein interaction resource. VirusMentha is an extension of a previous tool VirusMINT (Chatr-Aryamontri et al., 2009). VirusMentha is the most comprehensive viral PHI data source without limitation with respect to virus species or host organisms. The tool offers a graphical representation option for viral PHI networks (Calderone et al., 2015). On the other hand, ViRBase is a resource for virus–host ncRNA-associated interactions. It provides browsing and visualization of viral and cellular ncRNA-associated virus–virus, host–virus, and host–host interactions (Li et al., 2015).

Finally, the Web-based PHI databases comprising all pathogen types with known interactions are PHIDIAS (Xiang et al., 2007), HPIDB (Kumar and Nanduri, 2010), and PHISTO (Durmuş Tekir et al., 2013). PHIDIAS (Pathogen–Host Interaction Data Integration and Analysis System) stores data on genome sequences, conserved domains, and gene expression data related to PHIs. In addition to data storage, PHIDIAS offers the analysis of these data (Xiang et al., 2007). HPIDB (Host–Pathogen Interaction Database) is not limited to any pathogen or host regarding pathogen–host PPI data. HPIDB offers the BLASTP search option that allows searching for homologous PHI data for pathogens without experimental PHI data (Kumar and Nanduri, 2010). Currently, PHISTO (Pathogen-Host Interaction Search Tool) is the most comprehensive PHI database on the Web including data for all pathogenic microorganisms for which experimental protein interactions with human are available. Bioinformatic analysis tools in PHISTO allow users to visualize and analyze PHI networks to get insights on infection mechanisms (Durmuş Tekir et al., 2013). Using the tools in the current version of PHISTO, users can access the functional and topological properties of pathogen-targeted human proteins within the human intranetwork. Furthermore, a comparative analysis tool is provided to perform these analyses comparatively for different pathogens to observe the similarities and differences in their infection strategies.

Pathogen–host protein interaction data in the above PHI databases are integrated mainly from other PPI databases using automatic integration tools such as PSICQUIC (Aranda et al., 2011) and by manual curation from the literature. For the PHI tools, commonly used PPI databases including PHI data are

APID (Prieto and De Las Rivas, 2006), BIND (Alfarano et al., 2005), BioGrid (Chatr-aryamontri et al., 2013), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), IntAct (Orchard et al., 2013), iRefIndex (Razick et al., 2008), MINT (Licata et al., 2012), NetworKIN (Horn et al., 2014), Reactome (Croft et al., 2014), and STRING (Franceschini et al., 2013).

There are other informative databases for pathogens, providing useful information for studying infection mechanisms. For instance, ARDB (Antibiotic Resistance Genes Database) unifies most of the publicly available information on antibiotic resistance. The information can be used as a compendium of antibiotic resistance genes of newly sequenced genomes (Liu and Pop, 2009). IVDB (Influenza Virus Database) is an integrated information resource and analysis platform for influenza virus research focusing on the genetic, genomic, and phylogenetic studies. IVDB provides complete genome sequences of the virus to facilitate the analysis of global viral transmission and evolution (Chang et al., 2007). MPIDB (Microbial Protein Interaction Database) aims to collect all known physical interactions among the bacterial proteins (Goll et al., 2008). MvirDB is a microbial database of protein toxins, virulence factors, and antibiotic resistance genes for biodefense applications (Zhou et al., 2007). VFDB (Virulence Factor Database) is a comprehensive repository for bacterial virulence factors (Chen et al., 2011). VIDA is a virus database system for open reading frames (ORFs) of animal viruses (Albà et al., 2001). Finally, ViPR (Virus Pathogen Database and Analysis Resource) is an open bioinformatic resource for virology research. ViPR captures various types of information, including sequence data, gene, and protein annotations, 3D protein structures, clinical and surveillance metadata, and novel data derived from comparative genomics analyses (Pickett et al., 2012).

Text Mining of PHI Data from the Literature

Scientific publications are the main media through which researchers report their new findings. The huge amount and the continuing rapid growth of the number of published articles in biomedicine has made it particularly difficult for researchers to access and utilize the knowledge contained in them. Currently, there are over 24 million publications indexed in PubMed[11], which is the main system that provides access to the biomedical literature.

To address the challenge of information overload in the biomedical literature, a number of manually curated databases have been developed to store biologically important information such as protein interactions, gene–disease associations, or PHIs. However, given the current amount and the continuing rapid growth of the biomedical literature, it usually takes a lot of time and effort before new discoveries are included in these databases. Human database curation cannot keep up with literature production (Baumgartner et al., 2007). As a consequence, most of the knowledge remains hidden in the unstructured text of theh publised articles. Therefore, developing text mining techniques to uncover this knowledge has become an important research area. Several text mining approaches have been proposed for identifying articles relevant to a particular topic, detecting biomedical entities such as genes, proteins, and diseases in text, as well as extracting the relations among them. A number of shared tasks such as the BioCreative Challenges (Krallinger et al., 2008; Arighi et al., 2011) and the BioNLP Shared Tasks (Kim et al., 2009, 2011; Nédellec et al., 2013) have been conducted, which have further boosted research in this area. However, text mining for the pathogen-host interactions domain has not been well studied yet, although it has its own peculiarities and challenges. Only a handful of studies, which are discussed in the subsections below, have been conducted so far in this domain. One thread of research focuses on identifying the articles that contain PHI-relevant information (Yin et al., 2010; Korkin et al., 2011; Thieu et al., 2012) and another thread of research addresses performing more detailed semantic analysis of the text and extracting more fine-grained information such as the specific proteins that interact and the associated pathogen and host organisms (Korkin et al., 2011; Thieu et al., 2012).

PHI-Relevant Abstract Detection

Identifying and ranking articles that contain PHI-relevant information can be used for selecting and prioritizing articles for manual curation. It can also be an initial step for filtering the relevant articles before performing more fine-grained semantic analysis for identifying the biomedical entities and the relations among them. The task for detecting articles describing PPI information has been addressed in the BioCreative II, II.5, and III challenges (Krallinger et al., 2008; Leitner et al., 2010; Arighi et al., 2011). However, the focus has not been on PHI relevant articles. The first study that focused on detecting PHI-relevant abstracts, i.e., abstracts that describe pathogen host PPI, was conducted by (Yin et al., 2010). Similarly to most systems that participated in the BioCreative Challenges Article Classification Task, the problem was formulated as a supervised machine learning based classification task. Support Vector Machines (SVM) was used as the classification algorithm (Cortes and Vapnik, 1995). Feature selection methods including Information Gain, Mutual Information, and Chi-square were evaluated using a data set of 1360 manually labeled abstracts. The results showed that Information Gain and Chi-square perform better than Mutual Information as the number of features used decreases. Although the focus of the study was on PHI-relevant abstract classification, no any PHI specific features were used. Only the word unigrams and bigrams were used as features.

Pathogen–host interaction-relevant abstract classification was also tackled by (Thieu et al., 2012). Similarly to (Yin et al., 2010), the task was addressed as a supervised machine learning classification problem and SVM was used as the classification algorithm. However, unlike (Yin et al., 2010), the authors defined and used PHI specific features including the identified host and pathogen protein and gene names in the text, the host and pathogen organism names, the interaction signaling keywords, the experimental method keywords, and PHI-specific keywords such as virulence and effector. In order to account for the abstracts that report the absence of an interaction between a host and pathogen

[11] http://www.ncbi.nlm.nih.gov/pubmed

protein, features that make use of the negation signaling keywords were also designed. The protein and gene names, as well as the corresponding organisms were tagged by using the NLProt software (Mika and Rost, 2004). A set of dictionaries for interaction keywords, experimental keywords, negation keywords, PHI-keywords, host names, pathogen names, and uncertainty keywords was manually compiled. A data set of 175 PHI-relevant (positive set) and 175 PHI non-relevant (negative) abstracts was manually annotated and used for evaluation. The results showed that using PHI specific features is a promising approach for identifying PHI-relevant articles. However, it is not possible to compare the results with the results of (Yin et al., 2010), since a different data set was used for evaluation.

In order to be able to assess the performances of the proposed methods a larger and publicly available benchmark data set should be created. Such a data set should in fact contain three types of abstracts: (1) Abstracts that do not contain any PPI information (negative class 1); (2) Abstracts that contain PPI information which are not pathogen–host PPIs (negative class 2); and (3) Abstracts that contain pathogen–host PPI information (positive class). Distinguishing the positive class from negative class 2 is probably more difficult, since they both contain PPI information. The only difference is that the PPIs in negative class 2 are not PHIs. To distinguish these two classes from each other, PHI specific features should be utilized. On the other hand, distinguishing the positive class from negative class 1 is probably easier and generic PPI relevant features might be sufficient. It is not clear whether the data sets annotated and used in Yin et al. (2010) and Thieu et al. (2012) contain these three classes, or contain only two of them (i.e., the positive class and negative class 1). Therefore, it is difficult to assess and compare the reported results.

PHI-Relevant Relation Extraction

One of the most important opportunities for text mining in biomedicine is the identification of the relations among the biomolecules, which can help elucidate their roles in important biological processes, as well as in diseases. In order to extract the relations among biomedical entities from text, first the sequences of characters that correspond to entities should be tagged in text. This task is called Named Entity Recognition (NER) and has been an active research topic in the biomedical text mining domain.

While the earliest systems for biomedical NER were usually based on rule-based approaches (Fukuda et al., 1998), as annotated corpora became available, machine-learning based methods gained popularity (McDonald and Pereira, 2005; Tsai et al., 2006; Hsu et al., 2008). State-of-the-art gene and protein NER systems achieve a practically applicable level of performance (e.g., 87% *F*-score performance was obtained at the second BioCreative shared task on gene mention tagging (Smith et al., 2008)). Genia Tagger (Tsuruoka et al., 2005), ABNER (Settles, 2005), and BANNER (Leaman and Gonzalez, 2008) are some of the publicly available biomedical NER tools. LINNAEUS (Gerner et al., 2010) and OrganismTagger (Naderi et al., 2011) are tools developed for recognizing species names in biomedical text. Both achieve *F*-score performances of over 94%. Although the usability of these NER tools for the PHI domain has not been well addressed yet, in principle they can also be used for PHI text mining to identify the entity names such as gene, protein, and species names in text.

One of the first studies on using text mining for pathogen–host relationship extraction was conducted by (Anthony et al., 2010). As a case-study, the authors targeted the extraction of genotype, pathogen, and syndrome relations. A corpus consisting of 43 abstracts from PubMed was manually annotated. The available technologies for the automatic recognition of host–pathogen named entities and the relations among them were discussed. However, they have not been evaluated over the annotated corpus, which makes it difficult to draw conclusions about their usability for the PHI text mining domain.

Thieu et al. (2012) addressed the problem of extracting pathogen–host PPIs from text. The authors proposed a linguistically motivated approach that makes use of the link grammar representations of the sentences (Sleator and Temperley, 1995). Thieu et al. (2012) generated additional rules to map the protein names to the corresponding pathogen and host organism names. For instance, if an organism name occurs before a protein name (e.g., *Arabidopsis* RIN4 protein) the protein is mapped to the preceding organism. In addition, Thieu et al. (2012) incorporated an anaphora resolution module that resolves the pronouns such as "it," "they," etc. in the sentences with their corresponding protein/gene or organism names, which makes possible extracting relations that span multiple sentences. This module is based on the RelEx anaphora resolution method that uses the Hobbs' pronoun resolution algorithm (Hobbs, 1978). The proposed approach was evaluated by using the 350 annotated abstracts described in the section "PHI-Relevant Abstract Detection." The results of (Thieu et al., 2012) showed that the proposed approach significantly outperformed a naïve approach based on using one of the state-of-the-art generic PPI extraction tools Protein Interaction information Extraction (PIE) system (Kim et al., 2008). This motivates the development of methods that specifically address pathogen–host PPI extraction. The 24% *F*-score obtained by the proposed system suggests that there is room for improvement and further research in this domain is necessary. An error analysis suggested that an important source of error was the incorrect identification of protein names and incorrect assignment of species to the corresponding proteins. While the first one is a NER problem, which is an active research topic in biomedical text mining, the second one has not been tackled much by the researches. The results of the current studies suggest that it is a crucial research direction for PHI text mining studies.

Pathogen–host interaction-specific PPI extraction is a similar problem to the general problem of mining PPI relevant information from text (Ono et al., 2001; Blaschke and Valencia, 2002; Temkin and Gilder, 2003; Daraselia et al., 2004; Jelier et al., 2005; Erkan et al., 2007; Fundel et al., 2007; Airola et al., 2008; Tikk et al., 2010). However, it has its own peculiarities that require the development of methods specialized for PHI text mining. In order for a PPI to qualify as a PHI, the interaction should be intra-species. In other words, one of the proteins should be a host protein and the other one should be a pathogen protein. Therefore, besides tackling the problem of extracting the pair

of proteins that interact, the problems of identifying the species associated with them, as well as the classification of the species as host or pathogen should also be addressed. These additional requirements render the PHI text mining task more difficult than the already challenging PPI text mining task. Most PPI extraction systems operate on a sentence-level to extract the interactions. The underlying assumption is that the majority of the relations are contained within a single sentence. Analysis of the Genia event corpus (Kim et al., 2009) supports this assumption, since only 5% of the relations in the corpus span multiple sentences (Björne et al., 2009). However, this assumption does not in general hold for the PHI extraction task, since in many cases the species of the associated entities do not occur in the same sentence where the interaction is described (Thieu et al., 2012). Therefore, in order to extract PHIs from text, wider scope than a sentence should be considered and methods to merge information contained in multiple sentences should be developed. Nevertheless, the current findings from the generic PPI text mining domain can be utilized. For instance, recent studies have demonstrated the utility of integrating machine learning methods with similarity functions (or kernels) defined using the syntactic and semantic analysis of text (Tikk et al., 2010). Some of these approaches can be adapted to the PHI text mining domain by performing anaphora resolution as a prior step and extending the methods to operate on scopes wider than a sentence. In addition, novel methods should be developed to address the problem of assigning the species to their corresponding entities (e.g., proteins and genes). Sentence-level processing will probably not be sufficient to develop solutions to this problem, since species names do not necessarily occur in the same sentences or even in the same paragraphs as the entity names. Another challenge is that a species can be a host in one context, while it is a pathogen in another context. Therefore, methods for determining which species are pathogens and which are host in the given context should be designed.

The PHI information extracted using text mining can be utilized in at least two ways. First, such information can be used to populate PHI databases, either directly or indirectly by facilitating manual curation. This will make the data buried in the literature easily accessible to the researchers in this domain. Second, further analysis of the uncovered information can be integrated into a systems biology approach to generate new scientific hypothesis such as predicting currently unknown interactions among pathogen and host proteins.

Conclusion and Future Directions

Conventional therapeutics aim to kill pathogenic microorganisms directly usually by targeting the pathogen only. However, the drug resistance of pathogens demands alternative solutions for infectious threats, i.e., targeting host proteins required by pathogens for replication and persistence within the host organism or targeting PHIs (Murali et al., 2011; Zoraghi and Reiner, 2013). If these host proteins are indispensable for pathogens during infections, but not essential for host cells, they may serve as antimicrobial therapeutic targets to fight drug resistance. In parallel with the increase in the amount of PHI data, several genome-wide RNAi screening studies to identify cellular host factors were performed within the last decade (Ng et al., 2007; Brass et al., 2008; Hao et al., 2008; König et al., 2008, 2010; Krishnan et al., 2008; Zhou et al., 2008; Bushman et al., 2009; Li et al., 2009; Sessions et al., 2009; Tai et al., 2009; Karlas et al., 2010; Kumar et al., 2010; Murali et al., 2011; Moser et al., 2013; Lee et al., 2014). The detailed knowledge about mechanisms of the relationships between these host factors and their targeting pathogens is required urgently to develop new and more effective antimicrobial therapeutics, necessitating a computational systems biology approach to PHIs.

The computational modeling of networks of interacting genes, transcripts, proteins, and metabolites is of great importance in biomedical research to understand molecular mechanisms of PHIs. The high-throughput experimental detection of levels of biomolecules (gene transcripts, proteins, and metabolites) via omics approaches as well as the detection of PHIs via high-throughput experiments has generated comprehensive datasets. The presented review has provided a snapshot of recent developments in this area and a survey about databases that store such infection-specific data. Using text mining is necessary to extract the PHI-relevant data that are only available in the text of the huge amount of scientific literature. Although biomedical text mining is an active research area, there are only a limited number of studies focusing on extracting PHI information. The lack of a publicly available data set ('gold standard') makes it difficult to evaluate and compare the current approaches. Besides reviewing the current studies, we have also provided future directions for research including analyzing the usability of the already available biomedical text mining methods for the PHI text mining task, developing novel approaches addressing the peculiarities and challenges of the PHI domain, and creating publicly available benchmark data sets in order to provide a better assessment of the different methods. We have also covered studies on the bioinformatic analysis of three types (protein-based, regulatory, and metabolic) of PHI networks. The integrative analysis of the high-throughput omics experiments using modeling approaches will not only elucidate the mechanisms of infection, but will help in the discovery of potential therapeutic targets and drugs through selective identification of essential genes, proteins, and metabolites for the pathogen. Despite the recent efforts reviewed above, the use of systems biology approaches to investigate PHI systems is still in its infancy, mostly because of data scarcity. Ongoing studies in the field will lead to more complete PHI networks in the coming decade, improving the PHI-based solutions to infectious diseases.

Acknowledgments

AÖ was supported by Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme. RG was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre/Transregio 124 FungiNet (subprojects B3 and INF).

References

Ahmad, F., Babalola, O. O., and Tak, H. I. (2012). Potential of MALDI-TOF mass spectrometry as a rapid detection technique in plant pathology: identification of plant-associated microorganisms. *Anal. Bioanal. Chem.* 404, 1247–1255. doi: 10.1007/s00216-012-6091-7

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinform.* 9(Suppl. 11):S2. doi: 10.1186/1471-2105-9-S11-S2

Ako-Adjei, D., Fu, W., Wallin, C., Katz, K. S., Song, G., Darji, D., et al. (2015). HIV-1, human interaction database: current status and new features. *Nucleic Acids Res.* 43, D566–D570. doi: 10.1093/nar/gku1126

Albà, M. M., Lee, D., Pearl, F. M., Shepherd, A. J., Martin, N., Orengo, C. A., et al. (2001). VIDA: a virus database system for the organization of animal virus genome open reading frames. *Nucleic Acids Res.* 29, 133–136. doi: 10.1093/nar/29.1.133

Albrecht, D., Guthke, R., Brakhage, A. A., and Kniemeyer, O. (2010). Integrative analysis of the heat shock response in *Aspergillus fumigatus*. *BMC Genomics* 11:32. doi: 10.1186/1471-2164-11-32

Albrecht, D., Kniemeyer, O., Brakhage, A. A., Berth, M., and Guthke, R. (2007). Integration of transcriptome and proteome data from human-pathogenic fungi by using a data warehouse. *J. Integr. Bioinform* 4, 52.

Albrecht, D., Kniemeyer, O., Mech, F., Gunzer, M., Brakhage, A., and Guthke, R. (2011). On the way toward systems biology of *Aspergillus fumigatus* infection. *Int. J. Med. Microbiol.* 301, 453–459. doi: 10.1016/j.ijmm.2011.04.014

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., et al. (2005). The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* 33, D418–D424. doi: 10.1093/nar/gki051

Altwasser, R., Linde, J., Buyko, E., Hahn, U., and Guthke, R. (2012). Genome-wide scale-free network inference for *Candida albicans*. *Front. Microbiol.* 3:51. doi: 10.3389/fmicb.2012.00051

Anthony, S., Sintchenko, V., and Coiera, E. (2010). Text mining for discovery of host–pathogen interactions. *Infect. Dis. Inform.* 2010, 149–165. doi: 10.1007/978-1-4419-1327-2_7

Antunes, L. C. M., Arena, E. T., Menendez, A., Han, J., Ferreira, R. B., Buckner, M. M., et al. (2011). Impact of salmonella infection on host hormone metabolism revealed by metabolomics. *Infect. Immun.* 79, 1759–1769. doi: 10.1128/IAI.01373-10

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8, 528–529. doi: 10.1038/nmeth.1637

Arighi, C. N., Lu, Z., Krallinger, M., Cohen, K. B., Wilbur, W. J., Valencia, A., et al. (2011). Overview of the biocreative III workshop. *BMC Bioinform.* 12(Suppl. 8):S1. doi: 10.1186/1471-2105-12-S8-S1

Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G., and Perelson, A. S. (2009). Kineticsof influenza a virus infection in human. *J. Virol.* 80, 7509–7590.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and Di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Mol. Syst. Biol.* 122, 78.

Bartel, P. L., Roecklein, J. A., SenGupta, D., and Fields, S. (1996). A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* 12, 72–77. doi: 10.1038/ng0196-72

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 23, i41–i48. doi: 10.1093/bioinformatics/btm229

Bazzani, S., Hoppe, A., and Holzhütter, H.-G. (2012). Network-based assessment of the selectivity of metabolic drug targets in *Plasmodium falciparum* with respect to human liver metabolism. *BMC Syst. Biol.* 6:118. doi: 10.1186/1752-0509-6-118

Berglund, E. C., Nystedt, B., and Andersson, S. G. (2009). Computational resources in infectious disease: limitations and challenges. *PLoS Comput. Biol.* 5:e1000481. doi: 10.1371/journal.pcbi.1000481

Beste, D. J. V., Nöh, K., Niedenführ, S., Mendum, T. A., Hawkins, N. D., Ward, J. L., et al. (2013). 13C-flux spectral analysis of host-pathogen metabolism reveals a mixed diet for intracellular *Mycobacterium tuberculosis*. *Chem. Biol.* 20, 1012–1021. doi: 10.1016/j.chembiol.2013.06.012

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., and Salakoski, T. (2009). "Extracting complex biological events with rich graph-based feature sets," *in Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task,* Stroudsburg, PA, 10–18.

Blaschke, C., and Valencia, A. (2002). The frame-based module of the SUISEKI information extraction system. *IEEE Intell. Syst.* 17, 14–20.

Bleves, S., Dunger, I., Walter, M. C., Frangoulidis, D., Kastenmüller, G., Voulhoux, R., et al. (2014). HoPaCI-DB: host-*Pseudomonas* and *Coxiella* interaction database. *Nucleic Acids Res.* 42, D671–D676. doi: 10.1093/nar/gkt925

BonHomme, M., Wong, S., Carter, C., and Scarlata, S. (2003). The pH dependence of HIV-1 capsid assembly and its interaction with cyclophilin A. *Biophys. Chem.* 105, 67–77. doi: 10.1016/S0301-4622(03)00063-2

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., and Jamshidi, N. (2010). Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.* 6, 422. doi: 10.1038/msb.2010.68

Bordbar, A., and Palsson, B. O. (2012). Using the reconstructed genome-scale human metabolic network to study physiology and pathology. *J. Intern. Med.* 271, 131–141. doi: 10.1111/j.1365-2796.2011.02494.x

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., et al. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921–926. doi: 10.1126/science.1152725

Brazma, A., Krestyaninova, M., and Sarkans, U. (2006). Standards for systems biology. *Nat. Rev. Genet.* 7, 593–605. doi: 10.1038/nrg1922

Brown, J. R., Magid-Slav, M., Sanseau, P., and Rajpal, D. K. (2011). Computational biology approaches for selecting host–pathogen drug targets. *Drug Discov. Today* 16, 229–236. doi: 10.1016/j.drudis.2011.01.008

Brown, S. A., Palmer, K. L., and Whiteley, M. (2008). Revisiting the host as a growth medium. *Nat. Rev. Microbiol.* 6, 657–666. doi: 10.1038/nrmicro1955

Bröker, B. M., and van Belkum, A. (2011). Immune proteomics of *Staphylococcus aureus*. *Proteomics* 11, 3221–3231. doi: 10.1002/pmic.201100010

Bumann, D. (2009). System-level analysis of *Salmonella metabolism* during infection. *Curr. Opin. Microbiol.* 12, 559–567. doi: 10.1016/j.mib.2009.08.004

Bushman, F. D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T. L., et al. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5:e1000437. doi: 10.1371/journal.ppat.1000437

Cakir, T., and Khatibipour, M. J. (2014). Metabolic network discovery by top-down and bottom-up approaches and paths for reconciliation. *Front. Bioeng. Biotechnol.* 2:62. doi: 10.3389/fbioe.2014.00062

Calderone, A., Licata, L., and Cesareni, G. (2015). VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res.* 43, D588–D592. doi: 10.1093/nar/gku830

Calderwood, M. A., Venkatesan, K., Xing, L., Chase, M. R., Vazquez, A., Holthaus, A. M., et al. (2007). Epstein–Barr virus and virus human protein interaction maps. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7606–7611. doi: 10.1073/pnas.0702332104

Camilios-Neto, D., Bonato, P., Wassem, R., Tadra-Sfeir, M. Z., Brusamarello-Santos, L. C., Valdameri, G., et al. (2014). Dual RNA-seq transcriptional analysis of wheat roots colonized by *Azospirillum brasilense* reveals up-regulation of nutrient acquisition and cell cycle genes. *BMC Genomics* 15:378. doi: 10.1186/1471-2164-15-378

Cash, P. (2011). Investigating pathogen biology at the level of the proteome. *Proteomics* 11, 3190–3202. doi: 10.1002/pmic.201100029

Chang, S., Zhang, J., Liao, X., Zhu, X., Wang, D., Zhu, J., et al. (2007). Influenza virus database (IVDB): an integrated information resource and analysis platform for influenza virus research. *Nucleic Acids Res.* 35, D376–D380. doi: 10.1093/nar/gkl779

Chatr-Aryamontri, A., Breitkreutz, B.-J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–D823. doi: 10.1093/nar/gks1158

Chatr-Aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res.* 37, D669–D673. doi: 10.1093/nar/gkn739

Chavali, A. K., D'Auria, K. M., Hewlett, E. L., Pearson, R. D., and Papin, J. A. (2012). A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends Microbiol.* 20, 113–123. doi: 10.1016/j.tim.2011.12.004

Chen, L., Xiong, Z., Sun, L., Yang, J., and Jin, Q. (2011). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* 40, D641–D645. doi: 10.1093/nar/gkr989

Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y., Chen, R., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148, 1293–1307. doi: 10.1016/j.cell.2012.02.009

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Cujec, T. P., Cho, H., Maldonado, E., Meyer, J., Reinberg, D., and Peterlin, B. M. (1997). The human immunodeficiency virus transactivator Tat interacts with the RNA polymerase II holoenzyme. *Mol. Cell. Biol.* 17, 1817–1823.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* 20, 604–611. doi: 10.1093/bioinformatics/btg452

De Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., et al. (2008). Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230. doi: 10.1038/msb.2008.66

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103. doi: 10.1089/10665270252833208

Del Chierico, F., Petrucca, A., Vernocchi, P., Bracaglia, G., Fiscarelli, E., Bernaschi, P., et al. (2014). Proteomics boosts translational and clinical microbiology. *J. proteomics* 97, 69–87. doi: 10.1016/j.jprot.2013.10.013

Di Carli, M., Benvenuto, E., and Donini, M. (2012). Recent Insights into plant–virus interactions through proteomic analysis. *J. Proteome Res.* 11, 4765–4780. doi: 10.1021/pr300494e

Dolan, P. T., Zhang, C., Khadka, S., Arumugaswami, V., Vangeloff, A. D., Heaton, N. S., et al. (2013). Identification and comparative analysis of hepatitis C virus-host cell protein interactions. *Mol. Biosyst.* 9, 3199–3209. doi: 10.1039/c3mb70343f

Durmuş Tekir, S., Çakir, T., Ardiç, E., Sayilirbas, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen–host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Durmuş Tekir, S., Çakir, T., and Ülgen, K. Ö. (2012). Infection strategies of bacterial and viral pathogens through pathogen–human protein–protein interactions. *Front. Microbiol.* 3:46. doi: 10.3389/fmicb.2012.00046

Durmuş Tekir, S., and Ülgen, K. Ö. (2013). Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era. *Biotechnol. J.* 8, 85–96. doi: 10.1002/biot.201200110

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog.* 4:e32. doi: 10.1371/journal.ppat.0040032

Dyer, M. D., Neff, C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., et al. (2010). The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PLoS ONE* 5:e12089. doi: 10.1371/journal.pone.0012089

Eisenreich, W., Heesemann, J., Rudel, T., and Goebel, W. (2013). Metabolic host responses to infection by intracellular bacterial pathogens. *Front. Cell. Infect. Microbiol.* 3:24. doi: 10.3389/fcimb.2013.00024

Emmert-Streib, F., Dehmer, M., and Haibe-Kains, B. (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* 2:38. doi: 10.3389/fcell.2014.00038

Erkan, G., Özgür, A., and Radev, D. R. (2007). "Semi-supervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 228–237.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2008). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143. doi: 10.1038/nrmicro1949

Fields, S., and Song, O. (1989). A novel genetic system to detect protein protein interactions. *Nature* 340, 245–246. doi: 10.1038/340245a0

Finley, R. L., and Brent, R. (1994). Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc. Natl. Acad. Sci. U.S.A.* 91, 12980–12984. doi: 10.1073/pnas.91.26.12980

Fisher, S. K., Novak, J. E., and Agranoff, B. W. (2002). Inositol and higher inositol phosphates in neural tissues: homeostasis, metabolism and functional significance. *J. Neurochem.* 82, 736–754. doi: 10.1046/j.1471-4159.2002.01041.x

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspé, G., Tiollais, P., et al. (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242, 369–379. doi: 10.1016/S0378-1119(99)00511-9

Forsman, A., Rüetschi, U., Ekholm, J., and Rymo, L. (2008). Identification of intracellular proteins associated with the EBV-encoded nuclear antigen 5 using an efficient tap procedure and FT-ICR mass spectrometry. *J. Proteome Res.* 7, 2309–2319. doi: 10.1021/pr700769e

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

Fromont-Racine, M., Rain, J.-C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16, 277–282. doi: 10.1038/ng0797-277

Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). "Toward information extraction: identifying protein names from biological papers," in *proceedings of the Pacific Symposium on Biocomputing*, Hawaii, 707–718.

Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx—Relation extraction using dependency parse trees. *Bioinformatics* 23, 365–371. doi: 10.1093/bioinformatics/btl616

Gardiner, D. M., and Howlett, B. J. (2005). Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of *Aspergillus fumigatus*. *FEMS Microbiol. Lett.* 248, 241–248. doi: 10.1016/j.femsle.2005.05.046

Gardner, T. S., and Faith, J. J. (2005). Reverse-engineering transcription control networks. *Phys. Life Rev.* 2, 65–88. doi: 10.1016/j.plrev.2005.01.001

Gautier, V. W., Gu, L., O'Donoghue, N., Pennington, S., Sheehy, N., and Hall, W. W. (2009). In vitro nuclear interactome of the HIV-1 Tat protein. *Retrovirology* 6, 47. doi: 10.1186/1742-4690-6-47

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinform.* 11:85. doi: 10.1186/1471-2105-11-85

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics* 24, 1743–1744. doi: 10.1093/bioinformatics/btn285

Gottschalk, R. A., Martins, A. J., Sjoelund, V. H., Angermann, B. R., Lin, B., and Germain, R. N. (2013). Recent progress using systems biology approaches to better understand molecular mechanisms of immunity. *Semin. Immunol.* 25, 201–208. doi: 10.1016/j.smim.2012.11.002

Gottwein, E., and Cullen, B. R. (2008). Viral and cellular microRNAs as determinants of viral pathogenesis and immunity. *Cell host Microbe* 3, 375–387. doi: 10.1016/j.chom.2008.05.002

Gouzy, A., Poquet, Y., and Neyrolles, O. (2014). Nitrogen metabolism in *Mycobacterium tuberculosis* physiology and virulence. *Nat. Rev. Microbiol.* 12, 729–737. doi: 10.1038/nrmicro3349

Greenfield, A., Madar, A., Ostrer, H., and Bonneau, R. (2010). DREAM4: combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS ONE* 5:e13397. doi: 10.1371/journal.pone.0013397

Guillet, J., Hallier, M., and Felden, B. (2013). Emerging functions for the *Staphylococcus aureus* RNome. *PLoS Pathog.* 9:e1003767. doi: 10.1371/journal.ppat.1003767

Guirimand, T., Delmotte, S., and Navratil, V. (2015). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic Acids Res.* 43, D583–D587. doi: 10.1093/nar/gku1121

Guo, Y., Luo, J., Wang, J., Wang, Y., and Wu, R. (2011). How to compute which genes control drug resistance dynamics. *Drug Discov. Today* 16, 339–344. doi: 10.1016/j.drudis.2011.02.004

Guthke, R., Linde, J., Mech, F., and Figge, M. T. (2012). Systems biology of microbial infection. *Front. Microbiol.* 3:328. doi: 10.3389/fmicb.2012.00328

Guthke, R., Möller, U., Hoffmann, M., Thies, F., and Töpfer, S. (2005). Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics* 21, 1626–1634. doi: 10.1093/bioinformatics/bti226

Guttman, D. S., McHardy, A. C., and Schulze-Lefert, P. (2014). Microbial genome-enabled insights into plant-microorganism interactions. *Nat. Rev. Genet.* 15, 797–813. doi: 10.1038/nrg3748

Hao, L., Sakurai, A., Watanabe, T., Sorensen, E., Nidom, C. A., Newton, M. A., et al. (2008). *Drosophila* RNAi screen identifies host genes important for influenza virus replication. *Nature* 454, 890–893. doi: 10.1038/nature07151

Hartlova, A., Krocova, Z., Cerveny, L., and Stulik, J. (2011). A proteomic view of the host-pathogen interaction: the host perspective. *Proteomics* 11, 3212–3220. doi: 10.1002/pmic.201000767

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models - a review. *Biosystems* 96, 86–103. doi: 10.1016/j.biosystems.2008.12.004

Heilmann, C. J., Sorgo, A. G., and Klis, F. M. (2012). News from the fungal front: wall proteome dynamics and host-pathogen interplay. *PLoS Pathog.* 8:e1003050. doi: 10.1371/journal.ppat.1003050

Heroven, A. K., and Dersch, P. (2014). Coregulation of host-adapted metabolism and virulence by pathogenic yersiniae. *Front. Cell. Infect. Microbiol.* 4:146. doi: 10.3389/fcimb.2014.00146

Heyl, K. A., Klassert, T. E., Heinrich, A., Müller, M. M., Klaile, E., Dienemann, H., et al. (2014). Dectin-1 is expressed in human lung and mediates the proinflammatory immune response to nontypeable *Haemophilus influenzae*. *mBio* 5, e01492–e01414. doi: 10.1128/mBio.01492-14

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183. doi: 10.1038/415180a

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua* 44, 311–338. doi: 10.1016/0024-3841(78)90006-2

Horn, F., Heinekamp, T., Kniemeyer, O., Pollmächer, J., Valiante, V., and Brakhage, A. A. (2012). Systems biology of fungal infection. *Front. Microbiol.* 3:108. doi: 10.3389/fmicb.2012.00108

Horn, F., Rittweger, M., Taubert, J., Lysenko, A., Rawlings, C., and Guthke, R. (2014). Interactive exploration of integrated biological datasets using context-sensitive workflows. *Front. Genet.* 5:21. doi: 10.3389/fgene.2014.00021

Hsieh, T. Y., Matsumoto, M., Chou, H. C., Schneider, R., Hwang, S. B., Lee, A. S., et al. (1998). Hepatitis C virus core protein interacts with heterogeneous nuclear ribonucleoprotein K. *J. Biol. Chem.* 273, 17651–17659. doi: 10.1074/jbc.273.28.17651

Hsu, C.-N., Chang, Y.-M., Kuo, C.-J., Lin, Y.-S., Huang, H.-S., and Chung, I.-F. (2008). Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics* 24, i286–i294. doi: 10.1093/bioinformatics/btn183

Hünniger, K., Lehnert, T., Bieber, K., Martin, R., Figge, M. T., and Kurzai, O. (2014). A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. *PLoS Comput. Biol.* 10:e1003479. doi: 10.1371/journal.pcbi.1003479

Huthmacher, C., Hoppe, A., Bulik, S., and Holzhütter, H.-G. (2010). Antimalarial drug targets in *Plasmodium falciparum* predicted by stage-specific metabolic network analysis. *BMC Syst. Biol.* 4:120. doi: 10.1186/1752-0509-4-120

Isci, S., Dogan, H., Ozturk, C., and Otu, H. H. (2014). Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics* 30, 860–867. doi: 10.1093/bioinformatics/btt643

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., et al. (2000). Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. U.S.A.* 97, 1143–1147. doi: 10.1073/pnas.97.3.1143

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J. R., McGovern, K. E., Clarke, S. C., et al. (2012). Global landscape of HIV-human protein complexes. *Nature* 481, 365–370.

Jelier, R., Jenster, G., Dorssers, L. C. J., van der Eijk, C. C., van Mulligen, E. M., Mons, B., et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics* 21, 2049–2058. doi: 10.1093/bioinformatics/bti268

Kafsack, B. F., and Llinás, M. (2010). Eating at the table of another: metabolomics of host-parasite interactions. *Cell host Microbe* 7, 90–99. doi: 10.1016/j.chom.2010.01.008

Karlas, A., Machuy, N., Shin, Y., Pleissner, K.-P., Artarini, A., Heuer, D., et al. (2010). Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature* 463, 818–822. doi: 10.1038/nature08760

Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003). Advances in flux balance analysis. *Curr. Opin. Biotechnol.* 14, 491–496. doi: 10.1016/j.copbio.2003.08.001

Kentner, D., Martano, G., Callon, M., Chiquet, P., Brodmann, M., Burton, O., et al. (2014). *Shigella* reroutes host cell central metabolism to obtain high-flux nutrient supply for vigorous intracellular growth. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9929–9934. doi: 10.1073/pnas.1406694111

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database–2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Khadka, S., Vangeloff, A. D., Zhang, C., Siddavatam, P., Heaton, N. S., Wang, L., et al. (2011). A physical interaction network of dengue virus and human proteins. *Mol. Cell. Proteomics* 10:M111.012187. doi: 10.1074/mcp.M111.012187

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (2009). "Overview of BioNLP'09 shared task on event extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Tas,* Stroudsburg, PA, 1–9.

Kim, J.-D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., and Tsujii, J. (2011). "Overview of BioNLP shared task 2011," in *Proceedings of the BioNLP Shared Task 2011 Workshop*, Portland, OR, 1–6.

Kim, S., Shin, S.-Y., Lee, I.-H., Kim, S.-J., Sriram, R., and Zhang, B.-T. (2008). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.* 36, W411–W415. doi: 10.1093/nar/gkn281

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Komarova, A. V., Combredet, C., Meyniel-Schicklin, L., Chapelle, M., Caignard, G., Camadro, J.-M., et al. (2011). Proteomic analysis of virus-host interactions in an infectious context using recombinant viruses. *Mol. Cell. Proteomics* 10:M110.007443. doi: 10.1074/mcp.M110.007443

König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H.-H., Bhattacharyya, S., et al. (2010). Human host factors required for influenza virus replication. *Nature* 463, 813–817. doi: 10.1038/nature08699

König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Irelan, J. T., et al. (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 135, 49–60. doi: 10.1016/j.cell.2008.07.032

Korkin, D., Thieu, T., Joshi, S., and Warren, S. (2011). "Mining host-pathogen interactions," in *Systems and Computational Biology – Molecular and Cellular Experimental Systems*, ed. N.-S. Yang (Rijeka: InTech), 163–184. doi: 10.5772/22016

Kraibooj, K., Park, H.-R., Dahse, H.-M., Skerka, C., Voigt, K., and Figge, M. T. (2014). Virulent strain of *Lichtheimia corymbifera* shows increased phagocytosis by macrophages as revealed by automated microscopy image analysis. *Mycoses* 57(Suppl. 3), 56–66. doi: 10.1111/myc.12237

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., and Valencia, A. (2008). Overview ofthe protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.* 9(Suppl. 2), S4. doi: 10.1186/gb-2008-9-s2-s4

Krishnan, M. N., Ng, A., Sukumaran, B., Gilfoy, F. D., Uchil, P. D., Sultana, H., et al. (2008). RNA interference screen for human genes associated with West Nile virus infection. *Nature* 455, 242–245. doi: 10.1038/nature07207

Kumar, D., Nath, L., Kamal, M. A., Varshney, A., Jain, A., Singh, S., et al. (2010). Genome-wide analysis of the host intracellular network that regulates survival of *Mycobacterium tuberculosis*. *Cell* 140, 731–743. doi: 10.1016/j.cell.2010.02.012

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinform.* 11(Suppl. 6):S16. doi: 10.1186/1471-2105-11-S6-S16

Kwofie, S. K., Schaefer, U., Sundararajan, V. S., Bajic, V. B., and Christoffels, A. (2011). HCVpro: hepatitis C virus protein interaction database. *Infect. Genet. Evol.* 11, 1971–1977. doi: 10.1016/j.meegid.2011.09.001

Law, G. L., Korth, M. J., Benecke, A. G., and Katze, M. G. (2013). Systems virology: host-directed approaches to viral pathogenesis and drug targeting. *Nat. Rev. Microbiol.* 11, 455–466. doi: 10.1038/nrmicro3036

Leaman, R., and Gonzalez, G. (2008). "BANNER: an executable survey of advances in biomedical named entity recognition," in *Proceedings of the Pacific Symposium on Biocomputing*, Kohala Coast, 652–663.

Lee, A. S.-Y., Burdeinick-Kerr, R., and Whelan, S. P. J. (2014). A genome-wide small interfering RNA screen identifies host factors required for vesicular stomatitis virus infection. *J. Virol.* 88, 8355–8360. doi: 10.1128/JVI.00642-14

Leitner, F., Mardis, S. A., Krallinger, M., Cesareni, G., Hirschman, L. A., and Valencia, A. (2010). An overview of BioCreative II. 5. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 385–399. doi: 10.1109/TCBB.2010.61

Le Rouzic, E., Mousnier, A., Rustum, C., Stutz, F., Hallberg, E., Dargemont, C., et al. (2002). Docking of HIV-1 Vpr to the nuclear envelope is mediated by the interaction with the nucleoporin hCG1. *J. Biol. Chem.* 277, 45091–45098. doi: 10.1074/jbc.M207439200

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40, D857–D861. doi: 10.1093/nar/gkr930

Lima, T. B., Pinto, M. F. S., Ribeiro, S. M., de Lima, L. A., Viana, J. C., Gomes Júnior, N., et al. (2013). Bacterial resistance mechanism: what proteomics can elucidate. *FASEB J.* 7, 1291–1303. doi: 10.1096/fj.12-221127

Linde, J., Hortschansky, P., Fazius, E., Brakhage, A. A., Guthke, R., and Haas, H. (2012). Regulatory interactions for iron homeostasis in *Aspergillus fumigatus* inferred by a systems biology approach. *BMC Syst. Biol.* 6:6. doi: 10.1186/1752-0509-6-6

Linde, J., Schulze, S., Henkel, S. G., and Guthke, R. (2015). Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J.* 14, 346–378.

Linde, J., Wilson, D., Hube, B., and Guthke, R. (2010). Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. *BMC Syst. Biol.* 4:148.doi: 10.1186/1752-0509-4-148

Li, Q., Brass, A. L., Ng, A., Hu, Z., Xavier, R. J., Liang, T. J., et al. (2009). A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16410–16415. doi: 10.1073/pnas.0907439106

Li, Y., Wang, C., Miao, Z., Bi, X., Wu, D., Jin, N., et al. (2015). ViRBase: a resource for virus-host ncRNA-associated interactions. *Nucleic Acids Res.* 43, D578–D582. doi: 10.1093/nar/gku903

Li-Pook-Than, J., and Snyder, M. (2013). iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. *Chem. Biol.* 20, 660–666. doi: 10.1016/j.chembiol.2013.05.001

Liu, B., and Pop, M. (2009). ARDB–antibiotic resistance genes database. *Nucleic Acids Res.* 37, D443–D447. doi: 10.1093/nar/gkn656

Longo, A. V., Burrowes, P. A., and Zamudio, K. R. (2014). Genomic studies of disease-outcome in host–pathogen dynamics. *Integr. Comp. Biol.* 54, 427–438. doi: 10.1093/icb/icu073

Lu, W., Lo, S. Y., Chen, M., Wu, K. J., Fung, Y. K., and Ou, J. H. (1999). Activation of p53 tumor suppressor by hepatitis C virus core protein. *Virology* 264, 134–141. doi: 10.1006/viro.1999.9979

Lusic, M., Marcello, A., Cereseto, A., and Giacca, M. (2003). Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter. *EMBO J.* 22, 6550–6561. doi: 10.1093/emboj/cdg631

Ma, H., and Goryanin, I. (2008). Human metabolic network reconstruction and its impact on drug discovery and development. *Drug Discov. Today* 13, 402–408. doi: 10.1016/j.drudis.2008.02.002

Manchanda, H., Seidel, N., Krumbholz, A., Sauerbrei, A., Schmidtke, M., and Guthke, R. (2014). Within-host influenza dynamics: a small-scale mathematical modeling approach. *Biosystems* 118, 51–59. doi: 10.1016/j.biosystems.2014.02.004

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Matsumoto, M., Hsieh, T., Zhu, N., VanArsdale, T., Hwang, S. B., Jeng, K.-S., et al. (1997). Hepatitis C virus core protein interacts with the cytoplasmic tail of lymphotoxin-beta receptor. *J. Virol.* 71, 1301–1309.

McCraith, S., Holtzman, T., Moss, B., and Fields, S. (2000). Genome-wide analysis of vaccinia virus protein–protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* 97, 4879–4884. doi: 10.1073/pnas.080078197

McDonald, R., and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinform.* 6(Suppl. 1):S6. doi: 10.1186/1471-2105-6-S1-S6

Mech, F., Thywissen, A., Guthke, R., Brakhage, A. A., and Figge, M. T. (2011). Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS ONE* 6:e19591. doi: 10.1371/journal.pone.0019591

Mika, S., and Rost, B. (2004). Protein names precisely peeled off free text. *Bioinformatics* 20, i241–i247. doi: 10.1093/bioinformatics/bth904

Milenbachs, A. A., Brown, D. P., Moors, M., and Youngman, P. (1997). Carbon-source regulation of virulence gene expression in *Listeria monocytogenes*. *Mol. Microbiol.* 23, 1075–1085. doi: 10.1046/j.1365-2958.1997.2711634.x

Mooney, M., McWeeney, S., Canderan, G., and Sékaly, R.-P. (2013). A systems framework for vaccine design. *Curr. Opin. Immunol.* 25, 551–555. doi: 10.1016/j.coi.2013.09.014

Morens, D. M., Folkers, G. K., and Fauci, A. S. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature* 430, 242–249. doi: 10.1038/nature02759

Moser, L. A., Pollard, A. M., and Knoll, L. J. (2013). A Genome-wide siRNA screen to identify host factors necessary for growth of the parasite *Toxoplasma gondii*. *PLoS ONE* 8:e68129. doi: 10.1371/journal.pone.0068129

Mulder, N. J., Akinola, R. O., Mazandu, G. K., and Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Comput. Struct. Biotechnol. J.* 11, 1–10. doi: 10.1016/j.csbj.2014.08.006

Murali, T. M., Dyer, M. D., Badger, D., Tyler, B. M., and Katze, M. G. (2011). Network-based prediction and analysis of HIV dependency factors. *PLoS Comput. Biol.* 7:e1002164. doi: 10.1371/journal.pcbi.1002164

Naderi, N., Kappler, T., Baker, C. J. O., and Witte, R. (2011). Organism tagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics* 27, 2721–2729. doi: 10.1093/bioinformatics/btr452

Naji, S., Ambrus, G., Cimermančič, P., Reyes, J. R., Johnson, J. R., Filbrandt, R., et al. (2012). Host cell interactome of HIV-1 Rev includes RNA helicases involved in multiple facets of virus production. *Mol. Cell. Proteomics* 11:M111.015313. doi: 10.1074/mcp.M111.015313

Nakajima, N., and Akutsu, T. (2014). Network completion for static gene expression data. *Adv. Bioinform.* 2014, 382452. doi: 10.1155/2014/382452

Nédellec, C., Bossy, R., Kim, J.-D., Kim, J., Ohta, T., Pyysalo, S., et al. (2013). "Overview of bionlp shared task 2013," in *Proceedings of the BioNLP Shared Task 2013 Workshop*, Sofia, 1–7.

Ngo, H. T. T., Pham, L. V., Kim, J.-W., Lim, Y.-S., and Hwang, S. B. (2013). Modulation of mitogen-activated protein kinase-activated protein kinase 3 by hepatitis C virus core protein. *J. Virol.* 87, 5718–5731. doi: 10.1128/JVI.03353-12

Ng, T. I., Mo, H., Pilot-Matias, T., He, Y., Koev, G., Krishnan, P., et al. (2007). Identification of host genes involved in hepatitis C virus replication by small interfering RNA technology. *Hepatology* 45, 1413–1421. doi: 10.1002/hep.21608

Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320. doi: 10.1038/msb.2009.77

Olszewski, K. L., Morrisey, J. M., Wilinski, D., Burns, J. M., Vaidya, A. B., Rabinowitz, J. D., et al. (2009). Host-parasite interactions revealed by *Plasmodium falciparum* metabolomics. *Cell Host Microbe* 5, 191–199. doi: 10.1016/j.chom.2009.01.004

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics* 17, 155–161. doi: 10.1093/bioinformatics/17.2.155

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Otto, A., van Dijl, J. M., Hecker, M., and Becher, D. (2014). The *Staphylococcus aureus* proteome. *Int. J. Med. Microbiol.* 304, 110–120. doi: 10.1016/j.ijmm.2013.11.007

Owsianka, A. M., and Patel, A. H. (1999). Hepatitis C virus core protein interacts with a human DEAD box protein DDX3. *Virology* 257, 330–340. doi: 10.1006/viro.1999.9659

Palmer, A. C., and Kishony, R. (2013). Understanding, predicting and manipulating the genotypic evolution of antibiotic resistance. *Nat. Rev. Genet.* 14, 243–248. doi: 10.1038/nrg3351

Panayidou, S., Ioannidou, E., and Apidianakis, Y. (2014). Human pathogenic bacteria, fungi, and viruses in *Drosophila*: disease modeling, lessons, and shortcomings. *Virulence* 5, 253–269. doi: 10.4161/viru.27524

Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* 2, 28–36. doi: 10.1038/nri700

Pichlmair, A., Kandasamy, K., Alvisi, G., Mulhern, O., Sacco, R., Habjan, M., et al. (2012). Viral immune modulators perturb the human molecular network by common and unique strategies. *Nature* 487, 486–490. doi: 10.1038/nature11289

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40, D593–D598. doi: 10.1093/nar/gkr859

Pittman, K. J., Aliota, M. T., and Knoll, L. J. (2014). Dual transcriptional profiling of mice and *Toxoplasma gondii* during acute and chronic infection. *BMC Genomics* 15:806. doi: 10.1186/1471-2164-15-806

Pollmächer, J., and Figge, M. T. (2014). Agent-Based model of human alveoli predicts chemotactic signaling by epithelial cells during early *Aspergillus fumigatus* infection. *PLoS ONE* 9:e111630. doi: 10.1371/journal.pone.0111630

Prieto, C., and De Las Rivas, J. (2006). APID: agile protein interaction dataanalyzer. *Nucleic Acids Res.* 34, W298–W302. doi: 10.1093/nar/gkl128

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., et al. (2010). Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS ONE* 5:e9202. doi: 10.1371/journal.pone.0009202

Qian, C., and Cao, X. (2013). Regulation of Toll-like receptor signaling pathways in innate immune responses. *Ann. N. Y. Acad. Sci.* 1283, 67–74. doi: 10.1111/j.1749-6632.2012.06786.x

Raghunathan, A., Reed, J., Shin, S., Palsson, B., and Daefler, S. (2009). Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. *BMC Syst. Biol.* 3:38. doi: 10.1186/1752-0509-3-38

Rain, J.-C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., et al. (2001). The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215. doi: 10.1038/35051615

Ramachandra, S., Linde, J., Brock, M., Guthke, R., Hube, B., and Brunke, S. (2014). Regulatory networks controlling nitrogen sensing and uptake in *Candida albicans*. *PLoS ONE* 9:e92734. doi: 10.1371/journal.pone.0092734

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinform.* 9:405. doi: 10.1186/1471-2105-9-405

Rienksma, R. A., Suarez-Diez, M., Spina, L., Schaap, P. J., and Martins Dos Santos, V. A. P. (2014). Systems-level modeling of mycobacterial metabolism for the identification of new (multi-) drug targets. *Semin. Immunol.* 26, 610–622. doi: 10.1016/j.smim.2014.09.013

Rohmer, L., Hocquet, D., and Miller, S. I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol.* 19, 341–348. doi: 10.1016/j.tim.2011.04.003

Ruppin, E., Papin, J. A., de Figueiredo, L. F., and Schuster, S. (2010). Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.* 21, 502–510. doi: 10.1016/j.copbio.2010.07.002

Saayman, S., Ackley, A., Turner, A.-M. W., Famiglietti, M., Bosque, A., Clemson, M., et al. (2014). An HIV-encoded antisense long noncoding RNA epigenetically regulates viral transcription. *Mol. Ther.* 22, 1164–1175. doi: 10.1038/mt.2014.29

Saenz, R. A., Quinlivan, M., Elton, D., Macrae, S., Blunden, A. S., Mumford, J. A., et al. (2010). Dynamics of influenza virus infection and pathology. *J. Virol.* 84, 3974–3983. doi: 10.1128/JVI.02078-09

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi: 10.1093/nar/gkh086

Sarker, M., Talcott, C., and Galande, A. K. (2013). In silico systems biology approaches for the identification of antimicrobial targets. *Methods Mol. Biol.* 993, 13–30. doi: 10.1007/978-1-62703-342-8_2

Sasikaran, J., Ziemski, M., Zadora, P. K., Fleig, A., and Berg, I. A. (2014). Bacterial itaconate degradation promotes pathogenicity. *Nat. Chem. Biol.* 10, 371–377. doi: 10.1038/nchembio.1482

Scharf, D. H., Heinekamp, T., Remme, N., Hortschansky, P., Brakhage, A. A., and Hertweck, C. (2012). Biosynthesis and function of gliotoxin in *Aspergillus fumigatus*. *Appl. Microbiol. Biotechnol.* 93, 467–472. doi: 10.1007/s00253-011-3689-1

Schmidt, F., and Völker, U. (2011). Proteome analysis of host-pathogen interactions: investigation of pathogen responses to the host cell environment. *Proteomics* 11, 3203–3211. doi: 10.1002/pmic.201100158

Schulze, S., Henkel, S. G., Driesch, D., Guthke, R., and Linde, J. (2015). Computationalprediction of molecular pathogen-host interactions based on dual transcriptome data. *Front. Microbiol.* 6:65. doi: 10.3389/fmicb.2015.00065

Sessions, O. M., Barrows, N. J., Souza-Neto, J. A., Robinson, T. J., Hershey, C. L., Rodgers, M. A., et al. (2009). Discovery of insect and human dengue virus host factors. *Nature* 458, 1047–1050. doi: 10.1038/nature07967

Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* 21, 3191–3192. doi: 10.1093/bioinformatics/bti475

Shapira, S. D., Gat-Viks, I., Shum, B. O. V., Dricot, A., de Grace, M. M., Wu, L., et al. (2009). A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell* 139, 1255–1267. doi: 10.1016/j.cell.2009.12.018

Simon, S., Guthke, R., Kamradt, T., and Frey, O. (2012). Multivariate analysis of flow cytometric data using decision trees. *Front. Microbiol.* 3:114. doi: 10.3389/fmicb.2012.00114

Singh, H., Khan, A. A., and Dinner, A. R. (2014). Gene regulatory networks in the immune system. *Trends Immunol.* 35, 211–218. doi: 10.1016/j.it.2014.03.006

Singh, I., Tastan, O., and Klein-Seetharaman, J. (2010). "Comparison of virus interactions with human signal transduction pathways," in *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, Niagara Falls, NY, 17–24. doi: 10.1145/1854776.1854785

Six, A., Mariotti-Ferrandiz, M. E., Chaara, W., Magadan, S., Pham, H.-P., Lefranc, M.-P., et al. (2013). The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front. Immunol.* 4:413. doi: 10.3389/fimmu.2013.00413

Skalsky, R. L., and Cullen, B. R. (2010). Viruses, microRNAs, and host interactions. *Annu. Rev. Microbiol.* 64, 123–141. doi: 10.1146/annurev.micro.112408.134243

Sleator, D. D., and Temperley, D. (1995). Parsing English with a link grammar. *ArXiv Prepr. Cmp-lg/*9508004.

Smith, L., Tanabe, L. K., nee Ando, R. J., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., et al. (2008). Overview of BioCreative II gene mention recognition. *Genome Biol.* 9(Suppl. 2):S2. doi: 10.1186/gb-2008-9-s2-s2

Stanberry, L., Mias, G. I., Haynes, W., Higdon, R., Snyder, M., and Kolker, E. (2013). Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites* 3, 741–760. doi: 10.3390/metabo3030741

Stebbins, C. E. (2005). Structural microbiology at the pathogen-host interface. *Cell. Microbiol.* 7, 1227–1236. doi: 10.1111/j.1462-5822.2005.00564.x

Tai, A. W., Benita, Y., Peng, L. F., Kim, S.-S., Sakamoto, N., Xavier, R. J., et al. (2009). A functional genomic screen identifies cellular cofactors of hepatitis C virus replication. *Cell Host Microbe* 5, 298–307. doi: 10.1016/j.chom.2009.02.001

Temkin, J. M., and Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* 19, 2046–2053. doi: 10.1093/bioinformatics/btg279

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host–pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 28, 867–875. doi: 10.1093/bioinformatics/bts042

Tierney, L., Kuchler, K., Rizzetto, L., and Cavalieri, D. (2012a). Systems biology of host-fungus interactions: turning complexity into simplicity. *Curr. Opin. Microbiol.* 15, 440–446. doi: 10.1016/j.mib.2012.05.001

Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., et al. (2012b). An interspecies regulatory network inferred from simultaneous RNA-seq of

Candida albicans invading innate immune cells. *Front. Microbiol.* 3:85. doi: 10.3389/fmicb.2012.00085

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS Comput. Biol.* 6:e1000837. doi: 10.1371/journal.pcbi.1000837

Tripathi, L. P., Kataoka, C., Taguwa, S., Moriishi, K., Mori, Y., Matsuura, Y., et al. (2010). Network based analysis of hepatitis C virus core and NS4B protein interactions. *Mol. Biosyst.* 6, 2539–2553. doi: 10.1039/c0mb00103a

Tsai, R. T.-H., Sung, C.-L., Dai, H.-J., Hung, H.-C., Sung, T.-Y., and Hsu, W.-L. (2006). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinform.* 7(Suppl. 5):S11. doi: 10.1186/1471-2105-7-S5-S11

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., et al. (2005). "Developing a robust part-of-speech tagger for biomedical text," in *Advances in Informatics – 10th Panhellenic Conference on Informatics,* LNCS, Vol. 3746, eds P. Bozanis and E. N. Houstis (Berlin: Springer-Verlag), 382–392.

Ud-Dean, S. M. M., and Gunawan, R. (2014). Ensemble inference and inferability of gene regulatory networks. *PLoS ONE* 9:e103812. doi: 10.1371/journal.pone.0103812

Urban, M., Pant, R., Raghunath, A., Irvine, A. G., Pedro, H., and Hammond-Kosack, K. E. (2015). The Pathogen-host interactions database (PHI-base): additions and future developments. *Nucleic Acids Res.* 43, D645–D655. doi: 10.1093/nar/gku1165

van Someren, E. P., Wessels, L. F. A., Backer, E., and Reinders, M. J. T. (2002). Genetic network modeling. *Pharmacogenomics* 3, 507–525. doi: 10.1517/14622416.3.4.507

Vialás, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A., and Gil, C. (2009). Proteopathogen, a protein database for studying *Candida albicans*-host interaction. *Proteomics* 9, 4664–4668. doi: 10.1002/pmic.200900023

von Mentzer, A., Connor, T. R., Wieler, L. H., Semmler, T., Iguchi, A., Thomson, N. R., et al. (2014). Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat. Genet.* 46, 1321–1326. doi: 10.1038/ng.3145

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., et al. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* 287, 116–122. doi: 10.1126/science.287.5450.116

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Weber, M., Henkel, S. G., Vlaic, S., Guthke, R., van Zoelen, E. J., and Driesch, D. (2013). Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. *BMC Syst. Biol.* 7:1. doi: 10.1186/1752-0509-7-1

Wenk, M. R. (2006). Lipidomics of host–pathogen interactions. *FEBS Lett.* 580, 5541–5551. doi: 10.1016/j.febslet.2006.07.007

Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10, 618–630. doi: 10.1038/nrmicro2852

Wolstencroft, K., Owen, S., du Preez, F., Krebs, O., Mueller, W., Goble, C., et al. (2011). The SEEK: a platform for sharing data and models in systems biology. *Methods Enzymol.* 500, 629–655. doi: 10.1016/B978-0-12-385118-5.00029-3

Wu, W., Tran, K. C., Teng, M. N., Heesom, K. J., Matthews, D. A., Barr, J. N., et al. (2012). The interactome of the human respiratory syncytial virus NS1 protein highlights multiple effects on host cell biology. *J. Virol.* 86, 7777–7789. doi: 10.1128/JVI.00460-12

Xiang, Z., Tian, Y., and He, Y. (2007). PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol.* 8, R150 doi: 10.1186/gb-2007-8-7-r150

Xu, G., Strong, M. J., Lacey, M. R., Baribault, C., Flemington, E. K., and Taylor, C. M. (2014). RNA CoMPASS: a dual approach for pathogen and host transcriptome analysis of RNA-seq datasets. *PLoS ONE* 9:e89445. doi: 10.1371/journal.pone.0089445

Yang, H., Ke, Y., Wang, J., Tan, Y., Myeni, S. K., Li, D., et al. (2011). Insight into bacterial virulence mechanisms against host immune response via the *Yersinia pestis*-human protein-protein interaction network. *Infection Immun.* 79, 4413–4424. doi: 10.1128/IAI.05622-11

Yin, L., Xu, G., Torii, M., Niu, Z., Maisog, J. M., Wu, C., et al. (2010). Document classification for mining host pathogen protein–protein interactions. *Artif. Intell. Med.* 49, 155–160. doi: 10.1016/j.artmed.2010.04.003

Zheng, J., Sugrue, R. J., and Tang, K. (2011). Mass spectrometry based proteomic studies on viruses and hosts–a review. *Anal. Chim. Acta* 702, 149–159. doi: 10.1016/j.aca.2011.06.045

Zhou, C. E., Smith, J., Lam, M., Zemla, A., Dyer, M. D., and Slezak, T. (2007). MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* 35, D391–D394. doi: 10.1093/nar/gkl791

Zhou, H., Jin, J., and Wong, L. (2013). Progress in computational studies of host-pathogen interactions. *J. Bioinform. Comput. Biol.* 11, 1230001. doi: 10.1142/S0219720012300018

Zhou, H., Xu, M., Huang, Q., Gates, A. T., Zhang, X. D., Castle, J. C., et al. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe* 4, 495–504. doi: 10.1016/j.chom.2008.10.004

Zoraghi, R., and Reiner, N. E. (2013). Protein interaction networks as starting points to identify novel antimicrobial drug targets. *Curr. Opin. Microbiol.* 16, 566–572. doi: 10.1016/j.mib.2013.07.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 February 2015
doi: 10.3389/fmicb.2015.00065

Computational prediction of molecular pathogen-host interactions based on dual transcriptome data

Sylvie Schulze[1], Sebastian G. Henkel[2], Dominik Driesch[2], Reinhard Guthke[1] and Jörg Linde[1]*

[1] *Department of Systems Biology and Bioinformatics, Leibniz-Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute, Jena, Germany*
[2] *BioControl Jena GmbH, Jena, Germany*

Edited by:
Saliha Durmus, Gebze Technical University, Turkey

Reviewed by:
Ikbal Agah Ince, Wageningen University and Research Centrum, Netherlands
Kazim Yalcin Arga, Marmara University, Turkey

***Correspondence:**
Jörg Linde, Department of Systems Biology and Bioinformatics, Leibniz-Institute for Natural Product Research and Infection Biology – Hans-Knoell-Institute, Beutenbergstr. 11a, 07745 Jena, Germany
e-mail: joerg.linde@hki-jena.de

Inference of inter-species gene regulatory networks based on gene expression data is an important computational method to predict pathogen-host interactions (PHIs). Both the experimental setup and the nature of PHIs exhibit certain characteristics. First, besides an environmental change, the battle between pathogen and host leads to a constantly changing environment and thus complex gene expression patterns. Second, there might be a delay until one of the organisms reacts. Third, toward later time points only one organism may survive leading to missing gene expression data of the other organism. Here, we account for PHI characteristics by extending NetGenerator, a network inference tool that predicts gene regulatory networks from gene expression time series data. We tested multiple modeling scenarios regarding the stimuli functions of the interaction network based on a benchmark example. We show that modeling perturbation of a PHI network by multiple stimuli better represents the underlying biological phenomena. Furthermore, we utilized the benchmark example to test the influence of missing data points on the inference performance. Our results suggest that PHI network inference with missing data is possible, but we recommend to provide complete time series data. Finally, we extended the NetGenerator tool to incorporate gene- and time point specific variances, because complex PHIs may lead to high variance in expression data. Sample variances are directly considered in the objective function of NetGenerator and indirectly by testing the robustness of interactions based on variance dependent disturbance of gene expression values. We evaluated the method of variance incorporation on dual RNA sequencing (RNA-Seq) data of *Mus musculus* dendritic cells incubated with *Candida albicans* and proofed our method by predicting previously verified PHIs as robust interactions.

Keywords: network inference, NetGenerator, transcriptomics, dual RNA-Seq, microarrays, gene regulatory networks, inter-species interactions

1. INTRODUCTION

Organisms need to constantly adapt to environmental changes. On a molecular level, this is mediated by complex signaling cascades, which transmit the signal to cell nuclei. Transcription factors bind to their target genes, which consequently leads to a change in gene expression. This way, biological systems adapt to new environmental conditions.

In most cases underlying networks are unknown. This is especially interesting for interacting organisms, such as pathogens and host. Both the experimental setup and the nature of PHIs exhibit certain characteristics: (i) pathogen and host are in a battle leading to constantly changing conditions, (ii) a change in gene expression is triggered by new environmental conditions and the response of one organism might initiate faster or persist longer than the response of the other organism and (iii) two different organisms interact and eventually one survives which can lead to missing data time points.

The immune system of the host is permanently active to recognize and eliminate infectious microorganisms. As a first line of defense, components of the innate immune system such as the complement system, immune cells, and antimicrobial peptides recognize pathogen-associated molecular patterns (PAMPs). In contrast, pathogens developed many strategies to evade these mechanisms. They can shield microbe-associated cell surface proteins, mimic host surfaces or secrete proteases degrading host immune proteins (Zipfel et al., 2011). Nevertheless, the interaction with host cells is also important for pathogens, e.g., to acquire nutrients and to replicate (Casadevall and Pirofski, 2000).

The transcriptome of pathogen and host can be measured by physical separation of pathogen and host cells before RNA extraction. This enables RNA extraction from pathogen and host at different time points. For example, Oosthuizen et al. (2011) used separate pathogen and host microarrays to measure the transcriptome of *Aspergillus fumigatus* and human epithelial cells. The advantage of microarrays is, that they are cheap, processing of raw data is fast and well-established (Zhao et al., 2014). On the other hand, the recently developed RNA-Seq technology (Nagalakshmi et al., 2008) opened up the opportunity to study transcriptomes at a high level of accuracy and depth, also of non-model organisms. With the advent of dual RNA-Seq it became

possible to measure transcriptomes of multiple species simultaneously without physical separation of cells. A promising research field for application are infection processes of mammalian cells by pathogens (Westermann et al., 2012).

Network inference is a systems biology approach which aims to reverse engineer underlying interaction networks based on gene expression data (Hecker et al., 2009). To account for dynamics in the change of gene expression, some tools reconstruct gene regulatory networks (GRNs) based on gene expression time series data (Gustafsson et al., 2005; Guthke et al., 2005; Gupta et al., 2011; Vlaic et al., 2012). Predicted networks suggest interactions for experimental validation, but can also put experimental findings in a bigger context (Smet and Marchal, 2010). While numerous tools are applied to predict single-species networks, e.g., (Bansal et al., 2006; Bonneau et al., 2006; Linde et al., 2010; Altwasser et al., 2012), few inter-species approaches have been published.

NetGenerator, a tool to infer small scale GRNs (Guthke et al., 2005; Toepfer et al., 2007; Weber et al., 2013), has been successfully applied to predict single-species GRNs (Linde et al., 2012; Ramachandra et al., 2014). NetGenerator infers gene-regulatory networks from gene expression time series data. The interactions and their strength are identified by a heuristic structure search and parameter optimization. The resulting model is described by ordinary differential equations and can be displayed as a directed network graph as well as simulated. In a pioneering study, the applicability of NetGenerator to predict PHI networks has been demonstrated (Tierney et al., 2012). However, this publication focused on the specific biological example while the requirements for data processing and for the algorithm to a broader class of PHI experiments are not discussed extensively.

Hereafter, we discuss a variety of aspects for dual RNA-Seq data acquisition and processing. Furthermore, we describe the application of the extended NetGenerator version to infer an inter-species GRN based on dual RNA-Seq data. Even though we focus on the novel technique RNA-Seq, most parts of the described workflow can be applied to microarray data. We evaluate the impact of multiple input stimuli on the inference accuracy with NetGenerator based on a benchmark example. The extended NetGenerator version handles missing data values, which we demonstrate with the same benchmark example. We further extended the algorithm and its application to consider variances in replicated measurement data. This is directly embedded in the inference process and indirectly through a robustness analysis. We applied this method to a real dual RNA-Seq data set of murine dendritic cells infected with *C. albicans* published by Tierney et al. (2012).

2. RESULTS

2.1. DUAL RNA-SEQ DATA

2.1.1. Data acquisition

RNA-Seq requires a certain amount of input RNA often in a microgram range, which is practically difficult to extract. Furthermore, mRNA should be enriched to avoid sequencing data being dominated by structural RNAs (Tariq et al., 2011). Additionally, the experimental setup needs to ensure that enough mRNA of both organisms can be extracted to obtain an appropriate sequencing depth (**Figure 1A**). Westermann et al. (2012) discuss various important limitations for dual RNA-Seq. One aspect is that different genome sizes of pathogen and host lead to different amounts of cellular RNA. It is estimated that for instance only 1.5% of the human genome encodes proteins (International Human Genome Sequencing Consortium, 2001). For that reason, we suggest to estimate an appropriate sequencing depth for both organisms based on their transcriptome sizes and recommend a genome coverage of at least 10. Tools like featureCounts return transcriptome sizes based on given annotation files as side products (Liao et al., 2014).

Furthermore, the pathogen-host cell ratio of the experimental setup, also known as multiplicity of infection (MOI), has to be considered. A high MOI results in more pathogenic RNA, but may also lead to a faster and stronger host response and less clinical relevance.

The number of reads required to achieve a good genome coverage in both species has to be estimated in advance. The number of reads needs to be calculated for the least abundant species based on the intended fold coverage, transcriptome size and read length. The total number of reads can be estimated through the ratio of the amount of extracted pathogen and host RNA.

Furthermore, sequencing parameters need to be set taking into account transcriptome sizes and how closely related studied species are. Number of reads, read length, strand-specificity and single / paired end sequencing have a great impact on the number of ambiguously mapped reads. For instance, Yazawa et al. (2013) sequenced 100-base-pair single-end reads of the grass *Sorghum bicolor* and the pathogenic fungus *Bipolaris sorghicola*. Pittman et al. (2014) sequenced 100-base-pair paired-end reads of *M. musculus* and the parasite *Toxoplasma gondii*.

Finally, data time points have to be determined. A change of the transcriptional program triggered by a stimulus is usually strong at the start of the response. Thus, in best case the organism adapts and the degree of transcriptional change decreases. The temporal onset and duration of transcriptional response of pathogen and host can be very different. To detect both responses, RNA extraction time points need to be chosen carefully. Small-scale experiments should be carried out in advance to determine good data time points.

2.1.2. Dual RNA-Seq data processing

Preprocessing and analysis of sequencing data and the selection of candidate genes is an important step in advance of network inference (**Figure 1B**). The output of RNA-Seq are raw reads, of which low quality bases need to be trimmed [e.g., with trimmomatic (Bolger et al., 2014), btrim (Kong, 2011)]. Pathogen and host read data is separated *in silico* by aligning reads to the reference genomes (mapping). Engström et al. (2013) compare various available mapping tools and evaluate the conservative MapSplice, TopHat and STAR with comparatively low run time as favorable. From this point on, pathogen and host data are processed separately.

Tools like featureCounts (Liao et al., 2014) and htseq-Counts (Anders et al., 2014) calculate the number of reads mapped to a feature, e.g., an exon or gene, to determine gene expression levels. Subsequently, differential gene expression can be tested. Various tools [e.g., edgeR (Robinson et al., 2010), DESeq2 (Love

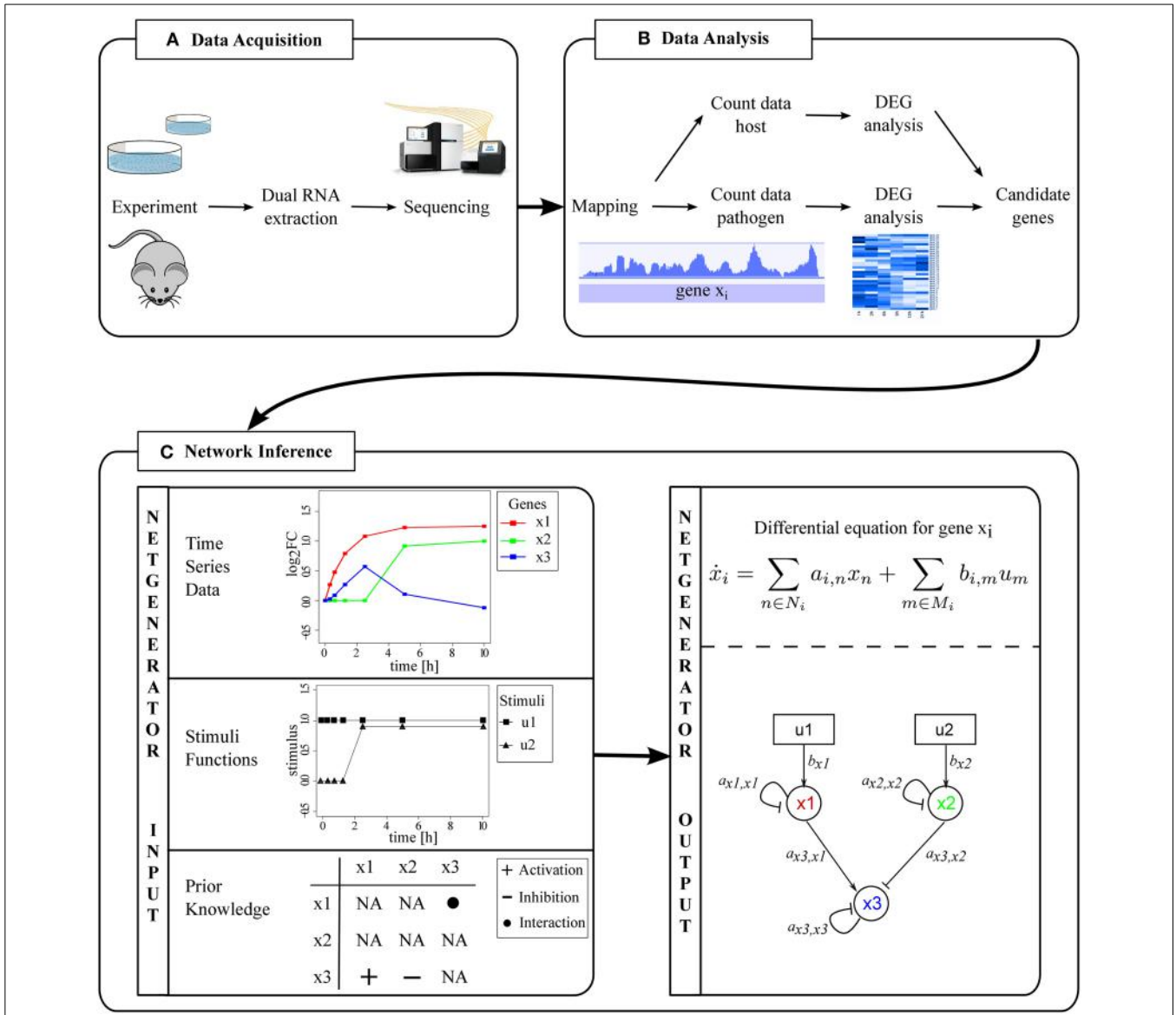


FIGURE 1 | From dual RNA-Seq data to inter-species GRNs. (A) Dual RNA extraction results in one sample to be sequenced. **(B)** Data preprocessing and analysis leads to separation of host and pathogen RNA-Seq data. DEGs are identified and candidate genes selected. **(C)** Prediction of an inter-species GRN with NetGenerator.

et al., 2014)] are available for that purpose and were reviewed recently (Soneson and Delorenzi, 2013; Zhang et al., 2014). The SEQC/MAQC-III Consortium recommends to apply pipeline dependent filters for *p*-value, fold change and expression-level to decrease estimated false discovery rates. Thereby, the outputs from different differential expression analysis pipelines yield a greater agreement (SEQC/MAQC-III Consortium, 2014).

Typically, hundreds of DEGs are found, of which a subset of candidate genes has to be selected. This number can be reduced, for instance by clustering gene expression kinetics (Bezdek, 1992) and choosing one representative for each cluster. This is advantageous, because it results in a set of candidate genes representing the major expression kinetics of the system. Furthermore, gene enrichment analysis can be carried out to select functional relevant candidate genes. FungiFun2 is one of the few enrichment tools for fungi and includes 298 strains from 240 species (Priebe et al., 2015). On the other hand, many enrichment tools exist for vertebrates. The underlying algorithms can be divided into three classes of which each shows certain advantages and drawbacks. It is also recommended to apply multiple tools (Huang et al., 2009; Tipney and Hunter, 2010).

2.2. MODELING PHI DATA

We extended the heuristic network inference tool NetGenerator (see Data and Methods) and its application to predict PHI networks. NetGenerator requires logarithmic fold changes (logFCs)

of gene expression time series data that can be obtained by various technologies, such as RNA-Seq or microarrays. Furthermore, the user of NetGenerator has to provide at least one input stimulus representing the external signal leading to a change in gene expression. Also, prior knowledge can be provided by the user to support the inference process (**Figure 1C**). It can be integrated in a compulsory ("fix") or soft ("flexible") way.

We generated a benchmark example to evaluate the influence of different stimuli and missing data on the inference performance (see Data and Methods). The benchmark comprised six data points of seven genes and two stimuli (**Figure 2A**). Prior knowledge data sets of two, four, six or eight interactions were randomly generated. We applied the extended NetGenerator version to infer GRNs based on the benchmark data set and each prior knowledge data set (soft integration). For small networks as the benchmark example the number of possible solutions was already very high. On sum, 63 edges (49 gene to gene interactions and 14 stimulus to gene interactions) and 2^{63} network topologies were possible not even including the interaction sign.

2.2.1. Multiple stimuli improve network inference

Multiple stimuli trigger responses in both pathogen and host during infection, such as the mutual stimulation of pathogen and host. This can be translated into at least two stimuli—the host stimulating the pathogen and *vice versa*. Weber et al. (2013) published the previous NetGenerator version V2.0 which can integrate multiple stimuli. We tested the influence of one or two stimuli on the performance of NetGenerator based on the benchmark example and each prior knowledge data set (**Figure 2B**).

First, only one constant stimulus (Test-1) set to a value of 1 was given. In a second test, an additional stimulus set to 0 until 30 min and set to 1 afterwards (Test-2) was given (Supplementary Table S1). We calculated mean values of F-measure (**Figure 2C**), sensitivity and specificity for every prior knowledge data set to

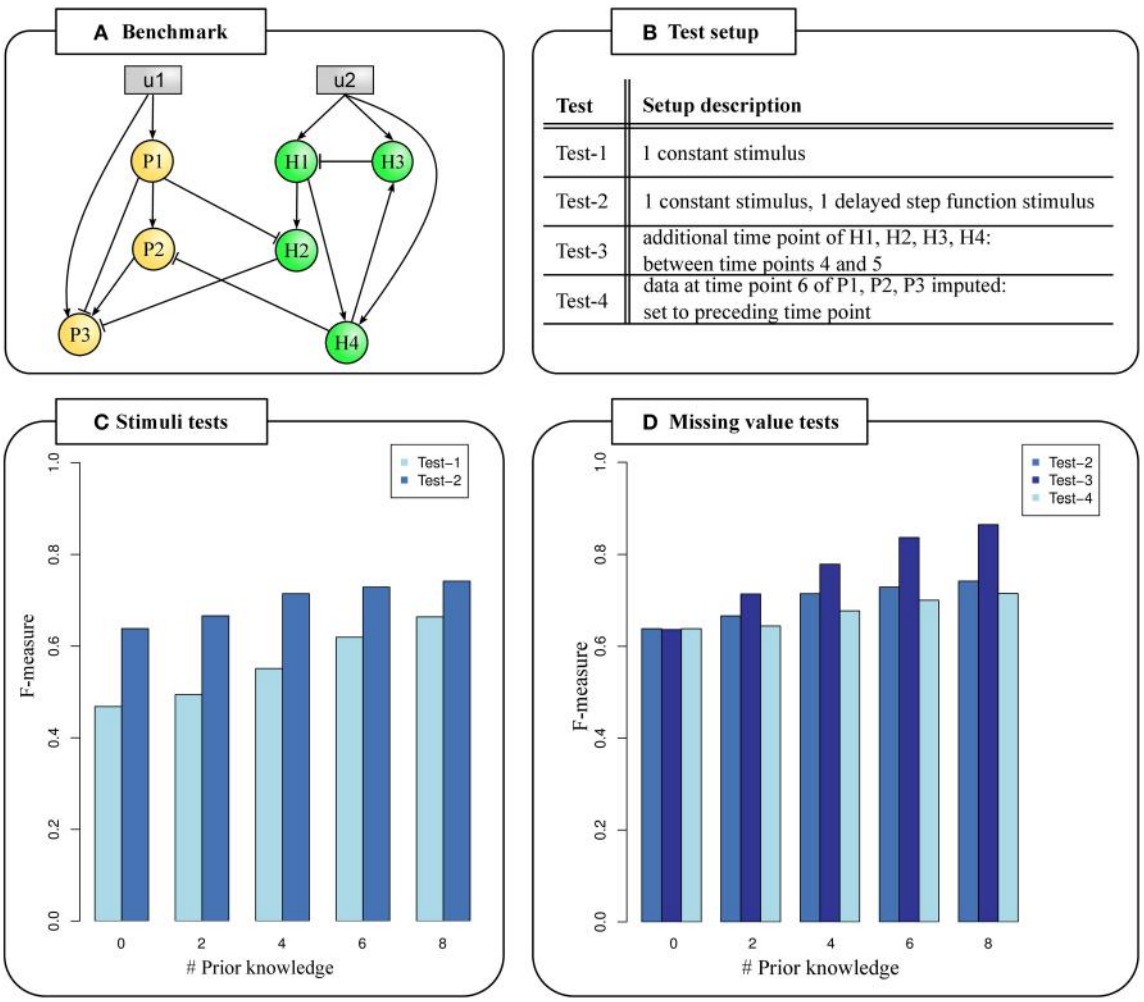


FIGURE 2 | Testing PHI data characteristics. (A) Benchmark example of an inter-species GRN with 3 pathogen candidate genes (orange nodes), four host candidate genes (green nodes) and two stimuli (gray nodes). Edges represent interactions. **(B)** Test setup. **(C)** F-measures calculated from predicted network topologies and the known network topology given different stimuli functions. Two stimuli increase F-measures (Test-2). **(D)** F-measures calculated from predicted network topologies and the known network topology based on missing data values. Carefully selected time points covering both the host and pathogen response increase F-measures (Test-3).

determine the accuracy of predicted GRNs in comparison to the known topology (Supplementary Table S2) (see Data and Methods).

We always observed noticeable larger F-measures given two stimuli in comparison to only one given stimulus. The difference in F-measure of Test-1 and Test-2 was up to 1.36 fold (**Figure 2C**). The less prior knowledge was given, the larger were the differences in F-measures between Test-1 and Test-2. We found the biggest performance difference between Test-1 and Test-2 when no or only two prior knowledge interactions were given. In these cases, 15 of 21 possible true positive edges were predicted when two stimuli were given, but only 11 true positive edges given one stimulus (Supplementary Table S2). In general, we observed increasing F-measures for more given prior knowledge independent of the number of stimuli.

2.2.2. Avoid missing data values

It is conceivable that time series experiments of pathogen and host were carried out independently under comparable experimental conditions. In this case, it is possible to utilize the pathogen and host data sets to predict PHI networks. Thus, data time points might differ which leads to missing values at intermediate time points or at the end of the time series. In case of dual RNA-Seq, pathogen and host are collectively processed. This may lead to a reduced amount of sample RNA of either of the species resulting in missing gene expression data. This is a problem especially for later time points when one species may dye. We extended the NetGenerator algorithm to handle missing data values at intermediate time points (see Data and Methods). We evaluated the influence of missing data on the performance based on the benchmark example, prior knowledge data sets and two given stimuli as in Test-2 (**Figure 2B**). Again, we calculated F-measure (**Figure 2D**), sensitivity and specificity (Supplementary Table S2).

We included data of one additional time point (165 min) for host genes, but additional data for pathogen genes were not given (Test-3). Thereby, we demonstrated the applicability of the extended NetGenerator version to data with missing values. We set the time point in such a way, that an additional data point covering the onset of the host response was provided and observed a noticeable increase of F-measure (**Figure 2D**). The difference in F-measure is greatest with 0.12 for eight given prior knowledge interactions. In this case, a mean number of 16.7 (Test-2) and 19.2 (Test-3) out of 21 possible true positive edges were predicted representing an improvement of 11.9%. This pointed out the importance of good time point selection covering both the pathogen and host response in a dual transcriptome data set.

NetGenerator requires complete data for the last time point. In case of missing measurements at the end of the time range for a subset of candidate genes, their values must be obtained in a different way and provided by the user. Here, we set the last time point to its preceding value (Test-4). We found slightly greater F-measures for Test-2 in comparison to Test-4 independent of the number of given prior knowledge. We observed a maximal difference between the F-measures between Test-2 and Test-4 (0.02) given four, six and eight prior knowledge interactions (**Figure 2D**).

2.3. INCORPORATION OF MEASUREMENT VARIANCES

Various differential expression analysis tools are available that calculate fold changes from multiple replicates. However, fold changes alone cannot reflect the degree of gene- and time point specific variances. This variance might be high especially regarding complex biological systems such as PHIs where cells from two species constantly interact and change the environment. However, biological variances can be considered in the network inference process to obtain robust predictions. For this purpose, we extended and applied NetGenerator to incorporate variances within the algorithm and in an outer robustness analysis. The extended NetGenerator algorithm was applied to one of the first published dual RNA-Seq data sets (Tierney et al., 2012) (see Data and Methods).

2.3.1. Extended NetGenerator algorithm incorporates measurement variances

Variances from replicated measurements were incorporated in the objective function of NetGenerator and need to be provided by the user. We calculated variances of the dual RNA-Seq data set of Tierney et al. (2012) as described (see Data and Methods).

We predicted a GRN (Supplementary Figure S1) with the extended NetGenerator based on logFCs and prior knowledge that were used as inputs for the previous NetGenerator in Tierney et al. (2012). Calculated gene- and time point specific variances were provided as input. Measured and simulated time courses of the GRN were plotted showing the standard deviations of measurements as shaded areas (**Figure 3A**). We observed that simulated data reproduced the measured data very well and were mostly within the shaded areas. Furthermore, simulated time courses were closer to data points with smaller standard deviation (e.g., *Hap3* at 30 min) than to data points with higher standard deviation (e.g., *Mta2* at 30 min).

2.3.2. Variance incorporation by an outer robustness analysis

Furthermore, variances were considered in an outer robustness analysis which we carried out based on the data of Tierney et al. (2012). The mean standard deviation was 1.24 with a minimum of 0.27 (*Sod5* at 120 min) and a maximum of 3.49 (*Mta2* at 30 min) (Supplementary Table S3). We scaled the standard deviations to a value of $\sigma_{\max} = 0.1$ (Supplementary Table S4). We calculated Gaussian distributed logFCs for every gene and time point (mean = measured logFC, σ = scaled standard deviation of replicates) (see Data and Methods). Thus, we generated 500 noisy data sets and applied the extended NetGenerator (**Figure 3B**). The robustness scores of the edges in the resulting 500 GRNs were illustrated in the bubble map (**Figure 3C**).

Tierney et al. (2012) experimentally verified the predicted inter-species interactions of *Ptx3* inhibiting *Hap3* and *Hap3* inhibiting *Mta2*. We predicted these verified interactions again as robust with the extended NetGenerator version (**Figure 3C**, Supplementary Table S5). Inhibition of *Hap3* by *Ptx3* was present in 71 % of predicted GRNs with a robustness score of 0.76. Inhibition of *Mta2* by *Hap3* was present in 72 % of predicted GRNs with a robustness score of 0.78. This also demonstrated the applicability of the presented robustness test.

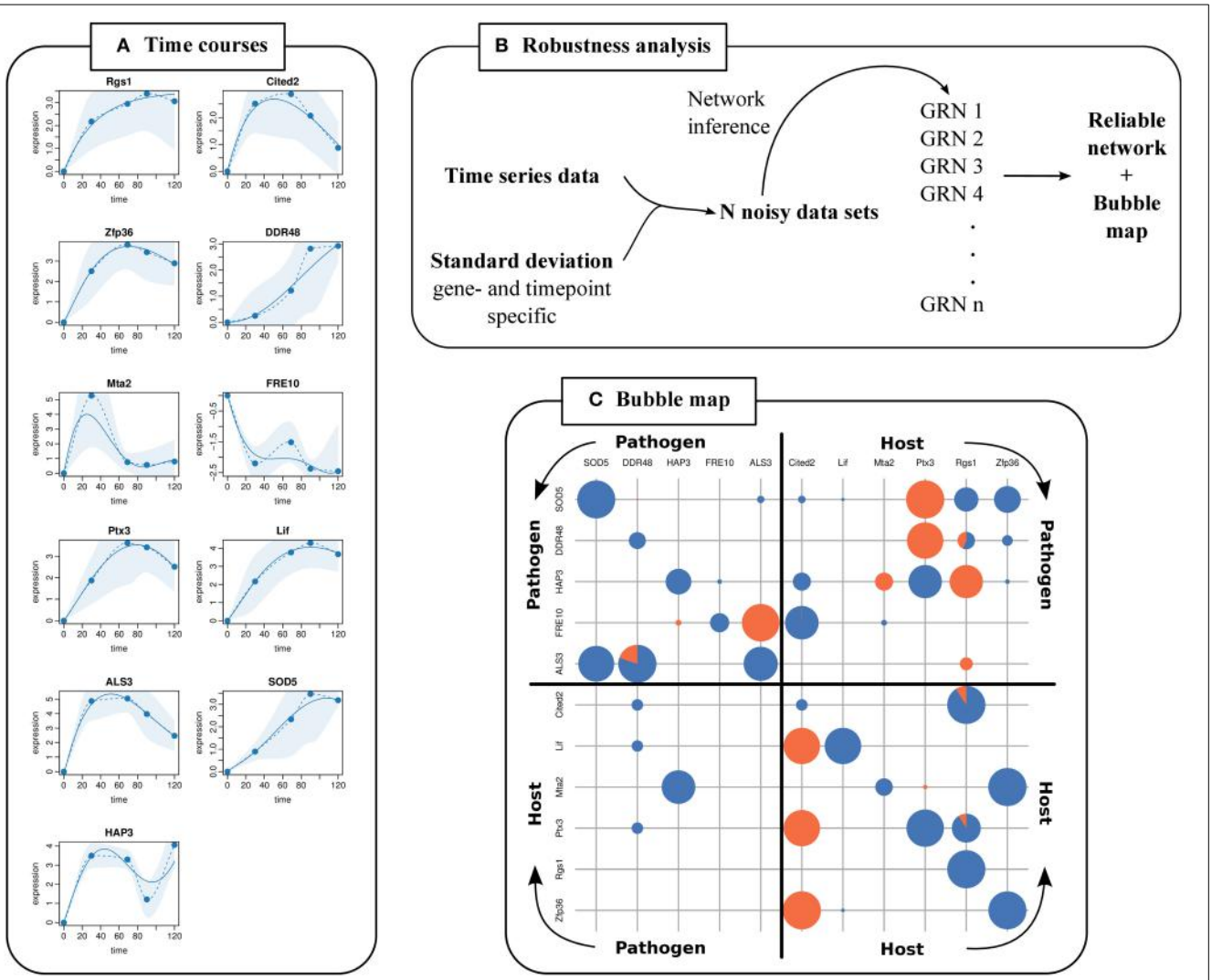


FIGURE 3 | GRN robustness analysis and visualization. (A) Fitting plots for each gene are generated showing measured time points (dots), simulated time courses (solid lines), interpolated time courses (dashed lines), and standard deviations from replicated measurements (shaded areas). **(B)** Outer robustness analysis. Noise is added to time series data with variances calculated from replicates of genes and time points. This is repeated *n* times to predict *n* GRNs. **(C)** The bubble map visualizes the robustness of a predicted edge from column gene to row gene. Bubble sizes illustrate the robustness score assigned to an edge. Orange and blue pies illustrate the fraction of activating and inhibiting edges, respectively.

3. DISCUSSION

In this study, we propose a workflow for dual RNA-Seq data acquisition, data processing and inter-species network inference. Furthermore, we describe how to handle a different temporal onset of transcriptional changes, missing data and how to integrate variances from replicated measurements based on the extended NetGenerator algorithm.

3.0.3. Delayed host response in PHI data

In a dual transcriptome data set we expect the onset of the pathogen and host transcriptional response at different time points. So far, several infection-related transcriptome studies of fungi were carried out. Transcriptome data was generated already at two to three time points within 60 min after infection (Linde et al., 2012; Ramachandra et al., 2014) suggesting an early onset of the pathogen's transcriptional response. This is further supported by a mechanism called adaptive prediction, that some pathogens have evolved. Based on cues from the current environment, pathogens predict a coming change in conditions and adapt their transcriptome in advance. An appropriate adaptation of the pathogen increases its survival chances (Brunke and Hube, 2014).

On the other hand, it takes some time until the host recognizes a pathogen. Moyes et al. (2010) showed that host epithelial cells initiate a response when a certain amount of pathogens exceeding a threshold is recognized. This is also a protective mechanism. Furthermore, the assumption of a later onset of the host transcriptional response is supported by various studies monitoring the host transcriptome from 1 h onwards (Banchereau et al., 2014; Favila et al., 2014).

However, we do not see a delayed transcriptional response of host DEGs in comparison to pathogen DEGs in the data set of Tierney et al. (2012) possibly because of the high MOI. Experimentalists keep improving their procedures to achieve realistic experimental setups, e.g., they decrease the MOI as much as possible still allowing them to extract the required amount of RNA for sequencing. Therefore, we expect to see a delayed host response in upcoming dual RNA-Seq data sets. To test the performance of the extended NetGenerator regarding different stimuli functions and missing data values, we generated a benchmark example showing a delayed onset of the host transcriptional response.

3.0.4. Gene expression time series data

NetGenerator requires time series gene expression data, at least one stimulus function and optionally prior knowledge. LogFCs are passed to NetGenerator in form of a data matrix, where columns correspond to candidate genes and rows to measured time points.

PHIs are very complex systems, but available data is limited regarding the number of time points and replicates. Furthermore, transcriptome data do not provide any information about processes taking place as for instance on protein level and in the extracellular space. Therefore, it has to be considered that predicted PHIs are indirect, when they are interpreted.

3.0.5. Modeling PHI stimuli

A GRN can be understood as a biological system that adapts to external, environmental stimuli yielding changes in gene expression. NetGenerator can integrate multiple stimuli and requires one function per stimulus representing it.

Many biological processes can be interpreted as external stimuli triggering responses in both pathogen and host cells during infection. In a typical experimental setup the host is incubated with the pathogen stimulating both organisms. The host recognizes PAMPs on pathogen cell surfaces by pathogen recognition receptors (PRRs). This initiates an information flow through signaling cascades (Akira et al., 2006). Nevertheless, the process of pathogen recognition resulting in a transcriptional response requires some time. Besides the molecular interaction with the host, the pathogen is also stimulated by different environmental factors, e.g., a change of temperature, pH and ion concentrations (Linde et al., 2010).

We found that multiple stimuli functions improve network inference results significantly. Therefore, we recommended to provide two or more stimuli functions for inter-species network inference. One option to model the stimulus representing the influence of the host on the pathogen is a constant function. Therewith, the stimulus is active from time point zero onwards and models an early pathogen transcriptional response. *Vice versa*, a second stimulus can represent the stimulation of the host by the pathogen. We predicted GRNs providing an additional input signal as a delayed step function (Test-2) aiming to model a later onset of the host transcriptional response. Another possible scenario would be to provide a stimulus function representing a slow increase of the influence.

More options for stimuli functions are possible when real experiments are carried out. For example, the number of differentially expressed host and pathogen genes can be determined for every time point and translated into stimuli functions. This can be done by scaling the number of DEGs to a range from zero to one. Additional measurements, e.g., cytokine release or cell contacts, can also be used as a basis for stimuli functions. Of particular interest is the growth curve of the pathogen, which we recommend to measure and integrate in the stimuli functions. Nevertheless, many biological events trigger responses, of which not all can be integrated in the network inference.

3.0.6. Prior knowledge sources

Optionally, the user of NetGenerator can provide prior knowledge about interactions of candidate genes. This is strongly recommended to reduce the search space resulting from the large number of possible interactions (Hecker et al., 2009). Prior knowledge can be softly integrated by assigning a score between zero and one that reflects its reliability. A score smaller than one allows prior knowledge to be rejected if it does not fit the data.

Prior knowledge about interactions in GRNs originates from published results that were transferred to databases. PHI databases like PHISTO (Tekir et al., 2013), PHI-base (Winnenburg et al., 2006), and HPIDB (Kumar and Nanduri, 2010) have been established. Mukherjee et al. (2013) listed various web sources of interaction data.

Host specific prior knowledge can be extracted manually from literature or automatically with text mining tools. Pathway Studio is a text mining tool specific for mammals (Nikitin et al., 2003). Further gene information is provided by organism specific websites, e.g., the human gene database GeneCards[1].

As well, organism specific websites exist for pathogens, e.g., Aspergillus Genome Database (Cerqueira et al., 2014) and Candida Genome Database (Inglis et al., 2012). To our knowledge, no fungi specific text mining tool is available. More general tools like GeneView—a semantic search engine for PubMed—can be applied (Thomas et al., 2012). Little is known about some pathogenic species. In this case, prior knowledge can be generated by searching orthologous genes in closely related and better studied organisms.

For both host and pathogen transcription factor binding motifs and binding sites can be obtained from databases, e.g., TRANSFAC (Matys et al., 2006), or predicted with bioinformatic tools as SiTaR (Fazius et al., 2011).

3.0.7. Robustness analysis

We extended the NetGenerator algorithm and its application to incorporate variances from replicated measurements in the inference method and in a robustness analysis. The output provides guidance for experimental validation of predicted interactions.

Inference methods should take into account the variance of replicates, because this additional information improves the parameter estimation. Under the assumption of independent Gaussian distributed noise the minimization of the objective function (Equation 4) corresponds to a Maximum Likelihood

[1] www.genecards.org

Estimator (MLE) (see e.g., Klipp et al., 2009, p. 155). Here, we assume that the variances of each gene and time point exhibit those statistical properties sufficiently. The extended NetGenerator version incorporates available measurement variances thus providing more reliable inference results. Nevertheless, the option to predict GRNs without providing variances is still available.

In previous publications a similar robustness analysis was carried out with the same standard deviation for each gene and time point set to a fixed value (Linde et al., 2010, 2012). Biological replicates can show high variance, that is gene- and time point specific and has a great influence on the estimated fold changes as well as their significance. Both the extended objective function (Equation 4) and the robustness analysis incorporate variances. They should be determined based on the available data to account for differences between genes and time points. One possibility is the rather simple approach to calculate the total variance of a logFC from sample variances as proposed (Equation 6). Another possibility is to derive the variances from software packages that take into account the statistical nature of the measurement method (including both biological and technical variances), perform processing steps, test for significant changes and determine logFCs. For instance, the R-package DESeq2 calculates standard errors for estimated logFCs (Love et al., 2014). Since those methods adjust the variances based on a statistical foundation, the inference results can be expected to further improve.

We performed the robustness analysis for the data of Tierney et al. (2012). In the data we observed very high variances for the replicates of some genes and time points. Applying the outer robustness analysis to noisy data sets based on unscaled standard deviations led to the prediction of diverse GRNs without more frequent edges. Therefore, we scaled the set of standard deviations to a maximal value. It is preferable to decrease the variance of expression mean by generating more biological replicates (Blainey et al., 2014).

The application of the robustness analysis is beneficial in many ways. It provides a ranking of predicted interactions based on noise added to the data. This makes it easier to decide, which predicted interactions should be experimentally verified. Furthermore, NetGenerator is a heuristic algorithm, which means that not all possible solutions are tested. It is likely, that not the best solution is returned, but a good one. The robustness analysis generates many good solutions resulting in a consensus network. It also accounts for possible mutually contradictory predictions.

4. DATA AND METHODS

4.1. APPLICATION OF EXTENDED NETGENERATOR TO PHI DATA

Network inference was carried out by the NetGenerator algorithm (see Guthke et al., 2005; Toepfer et al., 2007; Weber et al., 2013 for details). For this study, the previous NetGenerator V2.0 was extended (recent version of the R package: 2.3-0) to account for measurement variances and missing values.

4.1.1. Basic algorithm

The NetGenerator heuristics infers GRNs from time series gene expression data of multiple experiments and multiple stimulation. Expression data (logFCs), stimuli functions and prior knowledge (optionally) have to be provided by the user. Stimuli are factors that (directly or indirectly) cause changes in gene expression. It is assumed, that stimuli are not influenced by genes or their products, at least in the experimental setup. Nevertheless, stimuli values may evolve over time.

The inferred network model is described by a system of first order linear differential equations of the form

$$\dot{\underline{x}} = \underline{\underline{A}}\,\underline{x} + \underline{\underline{B}}\,\underline{u}. \tag{1}$$

The change of gene expression $\dot{\underline{x}}$ is influenced by other genes and (external) stimuli $\underline{u}$. While interactions between genes are described by the system matrix $\underline{\underline{A}} : N \times N$, the influence of stimuli is represented by the input matrix $\underline{\underline{B}} : N \times M$, where N is the number of genes and M is the number of inputs. The inference procedure determines the elements of these matrices, i.e., the parameters $\underline{\theta}$ of the model, by an iterative heuristics including structure and parameter optimization. In each iteration step, the algorithm includes a submodel which matches the available time series data best. The parameters of the ith submodel are determined by minimizing an objective function

$$J_i = J_{i,\text{output}} + J_{i,\text{priorknowledge}} \tag{2}$$

The second term evaluates the integration of prior knowledge, see (Weber et al., 2013) for details. In previous NetGenerator versions the first term

$$J_{i,\text{output}} = \sum_{e=1}^{E} \sum_{k=1}^{T_{e,i}} \left[w(t_k) \times \left(x_{e,i}(t_k) - \hat{x}_{e,i}(t_k, \underline{\theta}_i) \right)^2 \right] \tag{3}$$

described the error between measured data x and simulated data $\hat{x}$. The double sum was calculated for all experiments E and all time points $T_{e,i}$. Since the data contain both real and interpolated artificial values, this was accounted for by weighting factors $w(t_k)$.

4.1.2. Extension to account for missing values

NetGenerator was extended to account for missing data values. Now, NetGenerator accepts missing values at intermediate time points provided by the user as "NA." Internally, the time vector of the respective output is adjusted and interpolation is carried out based on existing measurement data. During inference, both simulation and objective function (Equation 4) can process that information of missing and replaced values.

4.1.3. Extension to incorporate variances

The objective function $J_{i,\text{output}}$ (Equation 3) was extended by additional weighting factors, which are the reciprocal variances $1/\sigma^2$ of the replicated data:

$$J_{i,\text{output}} = \sum_{e=1}^{E} \sum_{k=1}^{T_{e,i}} \left[\frac{w(t_k)}{\sigma_{e,i}^2(t_k)} \times \left(x_{e,i}(t_k) - \hat{x}_{e,i}(t_k, \underline{\theta}_i) \right)^2 \right] \tag{4}$$

Therefore, the variances σ^2 of the logFCs became additional input arguments to NetGenerator. Larger variances decrease the objective function value which effectively allows for a larger error between associated measured and simulated values in comparison to measurements of smaller variance.

4.1.4. Incorporation of variances in an outer robustness analysis

Moreover, variances are considered in an outer robustness analysis by predicting GRNs based on disturbed logFCs. To simulate the measurement process, we sampled three replicates of Gaussian distributed logFCs (mean = measured logFC , σ = standard deviation of replicates) and determined their mean. This resulted in a noisy logFC for each candidate gene and time point used as input for extended NetGenerator. We repeated this process 500 times.

For better visualization of the robustness analysis results we introduced the bubble map (**Figure 3C**) showing predicted interactions between candidate genes. It does not only consider the occurrence frequency of each edge, but also the sign and the respective objective function values $J = \sum J_i$ that is the sum over the values of each time series (Equation 2). The robustness score $S_{i,j}$ evaluating the interaction of gene j and gene i is calculated as

$$S_{i,j} = \sum_k \left\{ \frac{1}{J_{i,j,k}} \mid a_{i,j,k} \neq 0 \right\} \tag{5}$$

with $J_{i,j,k}$ being the objective function value of the k^{th} predicted GRN and $a_{i,j,k}$ being the corresponding element of the interaction matrix $\underline{\underline{A}}$. A robustness score $S_{i,j}$ of gene j interacting with gene i is illustrated by the bubble size of column j and row i (**Figure 3C**).

A big circle represents a frequently predicted interaction. Small or no circles represent rarely or no predicted interactions. Pie charts show the ratio of inferred activating (orange) and inhibiting (blue) interactions. Note, that the diagonal represents autoregulations. Exact robustness scores depending on how frequently an edge was predicted and corresponding objective function values of the predicted GRN are available as additional output (Supplementary Table S5).

4.1.5. Calculation of variances from replicates

Both the extended version of the objective function and the robustness analysis require variances derived from data. The gene- and time point specific variance σ_{tc}^2 of each logFC was calculated as the variance of the difference $\mu_t - \mu_c$ between means of treatment (t) and control (c) samples (error propagation):

$$\sigma_{tc}^2 = \sigma_c^2 + \sigma_t^2 \tag{6}$$

The respective standard deviations $\sigma_{i,j}$ of all genes and time points can be obtained by taking the square root of the variances. Given only few replicates, standard deviations can be high leading to the prediction of diverse GRNs. In that case, the standard deviations need to be scaled to a maximal value $\sigma_{\max}$:

$$\sigma_{i,j,\text{scaled}} = \frac{\sigma_{i,j} \times \sigma_{\max}}{max(\underline{\underline{\sigma}})} \tag{7}$$

4.2. DATA SETS AND EVALUATION CRITERIA

4.2.1. Benchmark model

We constructed a benchmark system composed of differential equations representing the logFC time series data of three pathogen genes, four host genes and two stimuli. The network topology included 21 directed, signed edges representing interactions. Common biological motifs like feed forward loops and feedback loops are integrated, too. Based on this topology we set up a system of differential equations and simulated this model with the R-package deSolve (Soetaert et al., 2010). We set the time point 0 min to zero and extracted data values of every differential equation at six time points on a logarithmic scale (15, 30, 60, 120, 250, 500 min). We added Gaussian distributed noise ($mean = 0$, $\sigma = 0.01$) to generate the benchmark data set.

As mentioned before, an additional input to guide network inference is prior knowledge. We generated a prior knowledge data set for the benchmark data by randomly sampling two interactions of the known network topology and repeated this 50 times. 50% of sampled prior knowledge is signed (activation or inhibition) and 50% is unspecific. Likewise, we generated prior knowledge data sets of four, six and eight interactions.

To evaluate predicted GRNs we computed statistical measures that compare the known topology to the predicted topology. Sensitivity (SE), specificity (SP), precision (PR) and F-measure (FM) are calculated as:

$$\begin{aligned} SE &= TP/(TP + FN + FP_s) \\ SP &= TN/(TN + FP_n) \\ PR &= TP/(TP + FP_n + FP_s) \\ FM &= (2 \times PR \times SE)/(PR + SE) \end{aligned} \tag{8}$$

taking the number of true positives (TP), false positives not part of the known topology (FP_n), false positives with wrong sign (FP_s), true negatives (TN) and false negatives (FN) into account (Weber et al., 2013). All of these statistical measures range from zero to one with one evaluating a predicted network as identical to the known topology.

4.2.2. Real dual RNA-Seq data

We utilized one of the first dual RNA-Seq data sets published by Tierney et al. (2012) as a second data set for evaluation. Murine dendritic cells were infected with *C. albicans* ($MOI = 5$). Three biological replicates were generated at 0, 30, 60, 90, 120 min after infection. Differential expression analysis was carried out with DESeq (Tierney et al., 2012). Six murine DEGs and five fungal DEGs were selected as candidate genes to predict an inter-species GRN with NetGenerator V1.0 (Toepfer et al., 2007). 19 prior knowledge edges were provided and softly integrated. We reproduced the result with NetGenerator V2.0 (Weber et al., 2013) based on the logFCs, stimulus function and prior knowledge of Tierney et al. (2012). Furthermore, we applied DESeq to determine logFCs and normalized count values to calculate gene- and time point specific variances.

The predicted interactions of *Ptx3* inhibiting *Hap3* and *Hap3* inhibiting *Mta2* were experimentally verified by Tierney et al. (2012). Therefore, these two interactions should

be again predicted by the extended NetGenerator and were thus used for evaluation.

4.2.3. Availability

The extended NetGenerator 2.3.-0 tool is available at `http://www.biocontrol-jena.com/NetGenerator/NetGenerator_2.3-0.tar.gz`.

AUTHOR CONTRIBUTIONS

Conception and design of the investigation and work: all. Analyzing the properties of PHIs: Sylvie Schulze, Jörg Linde, and Reinhard Guthke. Implementation of NetGenerator and contribution to mathematical background: Sebastian G. Henkel and Dominik Driesch. Data processing, application of computational algorithm and evaluation of results: Sylvie Schulze, Sebastian G. Henkel, and Jörg Linde. Drafting the manuscript: Sylvie Schulze and Sebastian G. Henkel. Revising it critically for important intellectual content and final approval of the version to be published: all. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: all.

FUNDING

Sylvie Schulze and Jörg Linde are supported by the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 "Pathogenic fungi and their human host: Networks of interaction," subproject B3 (Sylvie Schulze) and subproject INF (Jörg Linde). Sebastian G. Henkel and Dominik Driesch are supported within the Virtual Liver Network funded by the German Federal Ministry of Education and Research (BMBF, Fkz. 0315760).

ACKNOWLEDGEMENT

We would like to thank Uwe Menzel, Steffen Priebe and Sebastian Vlaic for valuable discussions about data processing and modeling.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00065/abstract

REFERENCES

Akira, S., Uematsu, S., and Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell* 124, 783–801. doi: 10.1016/j.cell.2006.02.015

Altwasser, R., Linde, J., Buyko, E., Hahn, U., and Guthke, R. (2012). Genome-wide scale-free network inference for *Candida albicans*. *Front. Microbiol.* 3:51. doi: 10.3389/fmicb.2012.00051

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq – a python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Banchereau, R., Baldwin, N., Cepika, A.-M., Athale, S., Xue, Y., Yu, C. I., et al. (2014). Transcriptional specialization of human dendritic Cell subsets in response to microbial vaccines. *Nat. Commun.* 5, 5283. doi: 10.1038/ncomms6283

Bansal, M., Gatta, G. D., and di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* 22, 815–822. doi: 10.1093/bioinformatics/btl003

Bezdek, J. C. (1992). *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*. New York, NY: Institute of Electrical and Electronics Engineers (IEEE) Press.

Blainey, P., Krzywinski, M., and Altman, N. (2014). Points of significance: replication. *Nat. Methods* 11, 879–880. doi: 10.1038/nmeth.3091

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., et al. (2006). The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.* 7:R36. doi: 10.1186/gb-2006-7-5-r36

Brunke, S., and Hube, B. (2014). Adaptive prediction as a strategy in microbial infections. *PLoS Pathog* 10:e1004356. doi: 10.1371/journal.ppat.1004356

Casadevall, A., and Pirofski, L. A. (2000). Host-Pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease. *Infect. Immun.* 68, 6511–6518. doi: 10.1128/IAI.68.12.6511-6518.2000

Cerqueira, G. C., Arnaud, M. B., Inglis, D. O., Skrzypek, M. S., Binkley, G., Simison, M., et al. (2014). The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.* 42, D705–D710. doi: 10.1093/nar/gkt1029

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., et al. (2013). Systematic evaluation of spliced alignment programs for RNA-Seq data. *Nat. Methods* 10, 1185–1191. doi: 10.1038/nmeth.2722

Favila, M. A., Geraci, N. S., Zeng, E., Harker, B., Condon, D., Cotton, R. N., et al. (2014). Human dendritic cells exhibit a pronounced type I IFN sigNature following *Leishmania major* infection that is required for IL-12 induction. *J. Immunol.* 192, 5863–5872. doi: 10.4049/jimmunol.1203230

Fazius, E., Shelest, V., and Shelest, E. (2011). SiTaR: a novel tool for transcription factor binding site prediction. *Bioinformatics* 27, 2806–2811. doi: 10.1093/bioinformatics/btr492

Gupta, R., Stincone, A., Antczak, P., Durant, S., Bicknell, R., Bikfalvi, A., et al. (2011). A computational framework for gene regulatory network inference that combines multiple methods and dataset. *BMC Syst. Biol.* 5:52. doi: 10.1186/1752-0509-5-52

Gustafsson, M., Hörnquist, M., and Lombardi, A. (2005). Constructing and analyzing a large-scale gene-to-gene regulatory network – lasso-constrained inference and biological validation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2, 254–261. doi: 10.1109/TCBB.2005.35

Guthke, R., Möller, U., Hoffmann, M., Thies, F., and Töpfer, S. (2005). Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics* 21, 1626–1634. doi: 10.1093/bioinformatics/bti226

Hecker, M., Lambeck, S., Töpfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models – a review. *Biosystems* 96, 86–103. doi: 10.1016/j.biosystems.2008.12.004

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Inglis, D. O., Arnaud, M. B., Binkley, J., Shah, P., Skrzypek, M. S., Wymore, F., et al. (2012). The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res.* 40, D667–D674. doi: 10.1093/nar/gkr945

International Human Genome Sequencing Consortium. (2001).Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2009). *Systems Biology: a Textbook*. Weinheim: Wiley-Blackwell.

Kong, Y. (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98, 152–153. doi: 10.1016/j.ygeno.2011.05.009

Kumar, R., and Nanduri, B. (2010). HPIDB - a unified resource for host-pathogen interactions. *BMC Bioinform.* 11(Suppl. 6), 16. doi: 10.1186/1471-2105-11-S6-S16

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Linde, J., Hortschansky, P., Fazius, E., Brakhage, A. A., Guthke, R., and Haas, H. (2012). Regulatory interactions for iron homeostasis in *Aspergillus fumigatus* inferred by a systems biology approach. *BMC Syst. Biol.* 6:6. doi: 10.1186/1752-0509-6-6

Linde, J., Wilson, D., Hube, B., and Guthke, R. (2010). Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. *BMC Syst. Biol.* 4:148. doi: 10.1186/1752-0509-4-148

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110. doi: 10.1093/nar/gkj143

Moyes, D. L., Runglall, M., Murciano, C., Shen, C., Nayar, D., Thavaraj, S., et al. (2010). A biphasic innate immune MAPK response discriminates between the yeast and hyphal forms of *Candida albicans* in epithelial cells. *Cell Host Microbe* 8, 225–235. doi: 10.1016/j.chom.2010.08.002

Mukherjee, S., Sambarey, A., Prashanthi, K., and Chandra, N. (2013). Current trends in modeling host-pathogen interactions. *Wiley Interdiscipl. Rev.* 3, 109–128. doi: 10.1002/widm.1085

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

Nikitin, A., Egorov, S., Daraselia, N., and Mazo, I. (2003). Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics* 19, 2155–2157. doi: 10.1093/bioinformatics/btg290

Oosthuizen, J. L., Gomez, P., Ruan, J., Hackett, T. L., Moore, M. M., Knight, D. A., et al. (2011). Dual organism transcriptomics of airway epithelial cells interacting with conidia of *Aspergillus fumigatus*. *PLoS ONE* 6:e20527. doi: 10.1371/journal.pone.0020527

Pittman, K. J., Aliota, M. T., and Knoll, L. J. (2014). Dual transcriptional profiling of mice and *Toxoplasma gondii* during acute and chronic infection. *BMC Genomics* 15:806. doi: 10.1186/1471-2164-15-806

Priebe, S., Kreisel, C., Horn, F., Guthke, R., and Linde, J. (2015). FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. *Bioinformatics* 31, 445–446. doi: 10.1093/bioinformatics/btu627

Ramachandra, S., Linde, J., Brock, M., Guthke, R., Hube, B., and Brunke, S. (2014). Regulatory networks controlling nitrogen sensing and uptake in *Candida albicans*. *PLoS ONE* 9:e92734. doi: 10.1371/journal.pone.0092734

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

SEQC/MAQC-III Consortium. (2014). A comprehensive assessment of RNA-Seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.* 32, 903–914. doi: 10.1038/nbt.2957

Smet, R. D., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717–729. doi: 10.1038/nrmicro2419

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving differential equations in R: package deSolve. *J. Stat. Softw.* 33, 1–25.

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-Seq data. *BMC Bioinform.* 14:91. doi: 10.1186/1471-2105-14-91

Tariq, M. A., Kim, H. J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNA-Seq analysis from minute amount of total RNA. *Nucleic Acids Res.* 39:e120. doi: 10.1093/nar/gkr547

Tekir, S. D., Çakır, T., Ardiç, E., Sayılırbaş, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen-host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., and Leser, U. (2012). GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res.* 40, W585–W591. doi: 10.1093/nar/gks563

Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., et al. (2012). An interspecies regulatory network inferred from simultaneous RNA-Seq of *Candida albicans* invading innate immune Cells. *Front. Microbiol.* 3:85. doi: 10.3389/fmicb.2012.00085

Tipney, H., and Hunter, L. (2010). An introduction to effective use of enrichment analysis software. *Hum. Genomics* 4, 202–206. doi: 10.1186/1479-7364-4-3-202

Toepfer, S., Guthke, R., Driesch, D., Woetzel, D., and Pfaff, M. (2007). "The NetGenerator algorithm: reconstruction of gene regulatory networks," in *Lecture Notes in Computer Science*, eds K. Tuyls, R. Westra, Y. Saeys, and A. Nowé (Berlin; Heidelberg: Springer), 119–130.

Vlaic, S., Schmidt-Heck, W., Matz-Soja, M., Marbach, E., Linde, J., Meyer-Baese, A., et al. (2012). The extended TILAR approach: a novel tool for dynamic modeling of the transcription factor network regulating the adaption to *in vitro* cultivation of murine hepatocytes. *BMC Syst. Biol.* 6:147. doi: 10.1186/1752-0509-6-147

Weber, M., Henkel, S. G., Vlaic, S., Guthke, R., van Zoelen, E. J., and Driesch, D. (2013). Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. *BMC Syst. Biol.* 7:1. doi: 10.1186/1752-0509-7-1

Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10, 618–630. doi: 10.1038/nrmicro2852

Winnenburg, R., Baldwin, T. K., Urban, M., Rawlings, C., Köhler, J., and Hammond-Kosack, K. E. (2006). PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res.* 34, D459–D464. doi: 10.1093/nar/gkj047

Yazawa, T., Kawahigashi, H., Matsumoto, T., and Mizuno, H. (2013). Simultaneous transcriptome analysis of sorghum and *Bipolaris sorghicola* by using RNA-Seq in combination with *de novo* transcriptome assembly. *PLoS ONE* 8:e62460. doi: 10.1371/journal.pone.0062460

Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., et al. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE* 9:e103207. doi: 10.1101/005611

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T Cells. *PLoS ONE* 9:e78644. doi: 10.1371/journal.pone.0078644

Zipfel, P. F., Skerka, C., Kupka, D., and Luo, S. (2011). Immune escape of the human facultative pathogenic yeast *Candida albicans*: the many faces of the *Candida* pra1 protein. *Int. J. Med. Microbiol.* 301, 423–430. doi: 10.1016/j.ijmm.2011.04.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 December 2014; accepted: 19 January 2015; published online: 06 February 2015.

Citation: Schulze S, Henkel SG, Driesch D, Guthke R and Linde J (2015) Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. Front. Microbiol. ***6****:65. doi: 10.3389/fmicb.2015.00065*

This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 20 July 2015
doi: 10.3389/fmicb.2015.00730

Reconstruction of the temporal signaling network in *Salmonella*-infected human cells

Gungor Budak[1], *Oyku Eren Ozsoy*[1], *Yesim Aydin Son*[1], *Tolga Can*[2] *and Nurcan Tuncbag*[1]*

[1] Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey,
[2] Department of Computer Engineering, College of Engineering, Middle East Technical University, Ankara, Turkey

Salmonella enterica is a bacterial pathogen that usually infects its host through food sources. Translocation of the pathogen proteins into the host cells leads to changes in the signaling mechanism either by activating or inhibiting the host proteins. Given that the bacterial infection modifies the response network of the host, a more coherent view of the underlying biological processes and the signaling networks can be obtained by using a network modeling approach based on the reverse engineering principles. In this work, we have used a published temporal phosphoproteomic dataset of *Salmonella*-infected human cells and reconstructed the temporal signaling network of the human host by integrating the interactome and the phosphoproteomic dataset. We have combined two well-established network modeling frameworks, the Prize-collecting Steiner Forest (PCSF) approach and the Integer Linear Programming (ILP) based edge inference approach. The resulting network conserves the information on temporality, direction of interactions, while revealing hidden entities in the signaling, such as the SNARE binding, mTOR signaling, immune response, cytoskeleton organization, and apoptosis pathways. Targets of the *Salmonella* effectors in the host cells such as CDC42, RHOA, 14-3-3δ, Syntaxin family, Oxysterol-binding proteins were included in the reconstructed signaling network although they were not present in the initial phosphoproteomic data. We believe that integrated approaches, such as the one presented here, have a high potential for the identification of clinical targets in infectious diseases, especially in the *Salmonella* infections.

Keywords: phosphoproteomic, network reconstruction, *Salmonella* infection, temporal data integration, pathway analysis

OPEN ACCESS

Edited by:
Saliha Durmus,
Gebze Technical University, Turkey

Reviewed by:
Jörg Linde,
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Germany
Tunahan Cakir,
Gebze Technical University, Turkey
Muhammed Erkan Karabekmez,
Boğaziçi University, Turkey

***Correspondence:**
Nurcan Tuncbag,
Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Universiteler Mah. Dumlupinar Blv. No. 1, Ankara 06800, Turkey
ntuncbag@metu.edu.tr

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *30 January 2015*
Accepted: *03 July 2015*
Published: *20 July 2015*

Citation:
Budak G, Eren Ozsoy O, Aydin Son Y, Can T and Tuncbag N (2015) Reconstruction of the temporal signaling network in Salmonella-infected human cells. Front. Microbiol. 6:730. doi: 10.3389/fmicb.2015.00730

Introduction

Salmonella enterica is a gastrointestinal pathogen that infects the human cells by translocation of its effector proteins with a vacuole called the *Salmonella* containing vacuole (SCV). The SCV compartment allows the pathogen to replicate and proliferate within the host cells. The secretion system transfers the pathogen proteins, which are called the effectors, directly into the cytosol of the host cells (reviewed in Dandekar et al., 2012). Translocation is achieved by type III secretion systems (T3SSs) where T3SS-1 is responsible for the regulation and replication of the SCV. These effectors have interactions with the host proteins and can change cell functions such as apoptosis, post-translational modifications, and intracellular signaling. *Salmonella* can

adapt to a broad range of environmental conditions and process many different metabolites (Dandekar et al., 2012). Although many efforts have been invested to understand the adaptation mechanism of *Salmonella*, functions of its effector proteins, the affected metabolic regulatory pathways, details of the host-pathogen communication and the changes in the host signaling pathways are still unknown. A systems level modeling approach has been performed on the effectors of *Salmonella* to understand the adaptation process and 14 regulators have been identified to play a critical role in the regulation of the genes responsible for *Salmonella* infection (Yoon et al., 2009). Also, the *Salmonella*'s metabolic network during its replication has been modeled using flux balance analysis, which has led to the identification of a set of metabolic pathways crucial during the intracellular replication (Raghunathan et al., 2009). The invasion of the pathogen is mainly transduced by the protein kinase signaling cascades in the host cell. *Salmonella* infection promotes apoptosis and adapts to the host cell's ubiquitination process (Steele-Mortimer, 2011). Regarding the regulome of *Salmonella*, the context likelihood of relatedness (CLR) approach has been used to infer the transcriptional regulatory connections by using mutual information in gene expression data and several regulatory networks have been identified (Taylor et al., 2009).

Understanding the communication and the signaling between *Salmonella* and its host in detail is crucial to improve the available treatment strategies for the *Salmonella* infection. The recently released interactome of the *Salmonella* effectors and human proteins, which has been curated from the literature, again revealed the enrichment of the MAPK signaling and the apoptotic pathways for the studied protein set (Schleker et al., 2012). The advances in high-throughput omic technologies also allow the systems-level identification of signaling components within the host cell. The analysis of mRNA expression of ∼4300 genes after a *Salmonella* infection in the human epithelial cells showed that NF-κB is a key transcription factor in the regulation of a wide range of genes (Eckmann et al., 2000). Also several cytokines, transcription factors and kinases are shown to be over-expressed in the same study. In a temporal gene expression analysis, where the Bayesian network analysis is used, the immune response, Wnt, PI3K, mTOR, TGF-β, and many other signaling pathways were found to be altered during the *Salmonella* infection. The host signal response was shown to be activated during the earlier time points rather than later (Lawhon et al., 2011). In another study, different gene expression datasets were integrated with protein–protein interactions and compared to each other to find out the specific subnetworks altered by *Salmonella* infection in the host (Dhal et al., 2014). In a global temporal phosphoproteomic analysis of *Salmonella*-infected human cells, 9500 phosphorylation events were quantified during the first 20 min of the infection and regulated host pathways were identified. Clustering analysis showed that the effector SopB was mainly responsible for the alterations of the phosphorylation events in the host cell (Rogers et al., 2011). Although omic technologies provide large amount of high dimensional data, the complete map of the signaling pathways cannot be retrieved by the direct connections of omic hits, as there are many intermediates which are not represented in the experimental data. Signaling networks can also be modeled by optimization based approaches (Dittrich et al., 2008; Huang and Fraenkel, 2009; Yeger-Lotem et al., 2009; Gosline et al., 2012; Huang et al., 2012; Tuncbag et al., 2013) where omic hits are defined as constraints. Previously, various network modeling approaches have been applied for the integration of the multiple omic sets of diseases and the disease networks of various cancer types have been successfully reconstructed (Kim et al., 2011; Huang et al., 2012). Regulatory networks can be reconstructed by various approaches, utilizing the gene expression data, including Boolean networks, Bayesian networks, and methods based on information theory and differential equations (reviewed in detail in De Jong, 2002; Hecker et al., 2009; Linde et al., 2015). Analysis of the perturbation data have also been proposed for the reconstruction networks (Markowetz et al., 2007; Frohlich et al., 2009; Bender et al., 2010; Aijo et al., 2013; Kiani and Kaderali, 2014). For example Nested Effects Models (NEMs) (Markowetz et al., 2007) use a set of knocked-down genes and their indirect effect on a larger set of genes to reconstruct the network. Methods that utilize observations of perturbed networks at a steady state or at several time points include (Dynamic) Deterministic Effects Propagation Networks [(D)DEPNs] (Frohlich et al., 2009; Bender et al., 2010), Sorad (Aijo et al., 2013), and Dynamic Probabilistic Boolean Threshold Networks (DPTBNs) (Kiani and Kaderali, 2014). However, these network reconstruction methods are computationally expensive and do not scale well for the reconstruction of large networks. Recently, Linear Programming (LP) based approaches have also been used to solve the network reconstruction problem (Eren Ozsoy and Can, 2013; Knapp and Kaderali, 2013; Matos et al., 2015). LP-based methods model the reconstruction problem as an optimization problem and are able to construct networks from both perturbation and time-series assays. However, based on the optimization function and the linear constraints, LP-based methods may be computationally expensive, as well. For example, a very recent method, lpNet (Matos et al., 2015), requires 3 days to reconstruct a 20 node network in the *in silico* dataset of the HPN-DREAM breast cancer network inference challenge. The DREAM (Dialogue on Reverse-Engineering Assessment and Methods) challenge aims to setup a joint effort between computational and experimental biologists toward revealing the cellular networks from multiple high-throughput data (Stolovitzky et al., 2007). An LP variant, the Integer Linear Programming (ILP) approach, by Melas et al. uses several optimization steps to find and remove the inconsistencies between measurements and the input network topology (Melas et al., 2013). Additionally, ILP is a known NP-hard problem, and due to the large number of variables, this method may not find solutions in a reasonable time, as it requires 64.000 s for a 14 node network. We have previously proposed a divide and conquer based ILP solution for the perturbation data analysis, which scales well for larger networks by merging the solutions of the smaller sub-networks (Eren Ozsoy and Can, 2013). The main difference of our proposed ILP approach was the definition of the optimization function as the minimization of the discrepancy between a reference network and the inferred network. Recently we have extended our ILP approach to work on time series data (Eren Ozsoy et al., 2015) and here we have directly apply that

method for the construction of the temporal signaling network in *Salmonella*-infected human cells. Although network modeling approaches are easily adaptable to identify signaling components in various disease states, to our best knowledge, these approaches have not yet been applied for the reconstruction of signaling networks in the human host cell during the *Salmonella* infection.

In this work, the temporal phosphoproteomic data of the *Salmonella*-infected human cells (Rogers et al., 2011) have been used to model the altered signaling network in the host cells. We have used a powerful combination of two different network modeling approaches to construct the signaling network of *Salmonella*-infected human cells. First, the temporal phosphoprotemic data of *Salmonella*-infected human cells are integrated with protein interaction data to construct the signaling pathway at each time point. Then, all constructed networks were merged together, and used as the input for the second part of the network modeling step, in which directions are assigned to the interactions based on the temporal data. Our approach allowed us to identify host pathways altered during *Salmonella* infection. In addition, by using network analysis techniques, we have provided a ranking of the proteins according to their importance during the infection.

Materials and Methods

Datasets

We have used the global temporal phoshoproteomic dataset published in (Rogers et al., 2011), where four time points, 2, 5, 10, and 20 min after *Salmonella* infection in human cell, were selected. Another dimension of this dataset is the cellular compartments where the phosphorylation site is identified as membrane, cytosol, or nucleus. At each time point, if the change in the phosphorylation status of a peptide is significantly altered compared to the uninfected cells ($p < 0.05$) and the variance across biological replicates are small (variance <15%) then that peptide is selected for the next step of the analysis, so added to the dataset. Then, each peptide selected, has been mapped to their HUGO Gene Nomenclature Committee (HGNC) identifiers using the Database for Annotation, Visualization and Integrated Discovery (DAVID) web server (Huang et al., 2009). If multiple peptides map to a single protein, then the peptide with the maximum value of fold change for phosphorylation level is included for further analysis.

Besides the global phoshoproteomic data, the human protein interactome is used for the data integration and modeling. The interaction data from iRefWeb has been downloaded which has 113,248 confident weighted interactions between 15,684 proteins (Turner et al., 2010). Also, a *Salmonella* effector to human host protein interactome, which consists of 40 effectors and 50 host proteins connected with 62 interactions, has been retrieved (Schleker et al., 2012).

Network Modeling

The network modeling procedure is composed of two stages; (i) network construction using the Prize-collecting Steiner Forest (PCSF) approach, and (ii) network reconstruction using the ILP based edge inference approach. These two approaches complement each other as the PCSF approach reveals the hidden components in signaling by finding the high confidence regions in the interactome, and the ILP-based edge inference approach reconstructs interactions and their directionality by using temporal data as constraints. In **Figure 1**, the flowchart of our integrated approach is given.

Prize-collecting Steiner Forest Approach

PCSF is based on finding the high-confidence regions within a protein interactome, which is used to recover the phosphoproteomic hits (i.e., the terminal nodes) and hidden proteins from the global temporal phoshoproteomic data (i.e., the Steiner nodes) in this study (Tuncbag et al., 2013). In the optimization stage, two objectives are important; avoiding the low-confidence protein-protein interactions in the final network and including as many phosphoproteomic hits as possible. Each protein identified with a significantly changed phosphorylation status at a time point is assigned a weight equal to the absolute value of the log fold change in phosphorylation, which is called the "prize." The cost of an interaction is calculated from its confidence score where higher cost implies lower confidence. The algorithm has to assign a cost for each interaction included in the network, and pay a penalty for excluding a phosphoproteomic hit equal to its prize. For a given, directed or undirected network G(V, E, $c(e)$, $p(v)$) with a node set V and edge set E, $p(v) \geq 0$ assigns a prize to each node $v \in$ V and $c(e) > 0$ assigns a cost to each edge $e \in$ E. The aim is to find a forest F(V_F,E_F) that minimizes the objective function:

$$f(t) = \sum\nolimits_{v \notin F} \left(\beta \cdot p(v)\right) + \sum\nolimits_{e \in F} \left(c(e)\right) + \omega \cdot \kappa \qquad (1)$$

where κ is the number of trees in the forest and β is the scaling factor. Another parameter that is used at the optimization stage is the depth value (D) which represents the maximum allowed number of edges from the root node to any terminal node. To convert the PCSF problem into a Prize-collecting Steiner Tree (PCST) problem we have introduced an extra root node v_0 into the network connected to each terminal node $t \in$ T by an edge (t,v_0) with cost ω where $\mathbf{T} \subset \mathbf{V}$. This optimization problem has been solved with a message passing algorithm, the msgsteiner tool (Bailly-Bechet et al., 2010). The forest F is defined as a disjoint collection of trees with all edges pointing to the roots. In this work, depth is set to 10, ω is in the interval [1, 10] and β is in the interval [1, 10]. Optimum forests obtained by each parameter combination are merged together in order to consider the suboptimal solutions. Finally, a PCSF is constructed for each time point.

Integer Linear Programming (ILP) Based Edge Inference Approach

For constructing signaling and regulatory networks using time series expression data, we have used an extended version of our previous ILP model that can handle both the time series and steady state perturbation data (Eren Ozsoy and Can, 2013). The extended ILP-based edge inference approach proposed in Eren Ozsoy et al. (2015) is highly scalable when time series data is available; therefore, in this paper, we directly apply that method

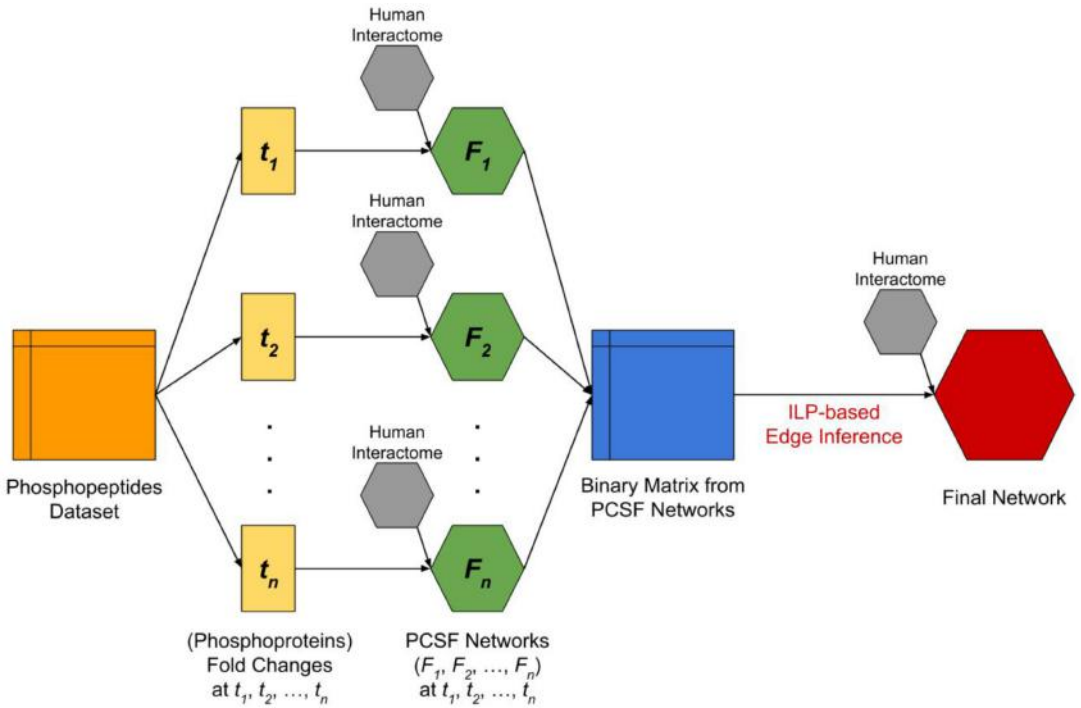


FIGURE 1 | The flowchart of complete analysis. The dataset which includes temporal fold changes of phosphopeptides at four different time points ($t_1 = 2$ min, $t_2 = 5$ min, $t_3 = 10$ min, $t_4 = 20$ min) and at three different locations (nucleus, cytoplasm, and membrane) was split and converted into temporal fold changes datasets of the corresponding phosphoproteins by taking the maximum fold change among phosphopeptides that were observed at different locations and mapped to the same phosphoprotein. Next, we applied PCSF approach for each fold changes dataset by integrating human interactome in order to discover hidden intermediate proteins. The resulting networks (F_1, F_2, F_3, F_4) are then used to form a binary matrix where the rows are time points and columns are phosphoproteins. Each corresponding cell of the binary matrix represents a significant change ($p < 0.05$ and variance <15%) in the phosphoprotein at the time point. Finally, we applied an ILP-based edge inference approach by integrating human interactome in order to validate and determine edges and edge directions.

for the construction of the whole *Salmonella* infection signaling network using a single integer linear program. The details of the ILP model are given below.

The integer linear programming model

Assuming that a reference signaling network is given as a directed graph G(V, E), where V represents the node set (i.e., proteins) and E represents the edge set (i.e., pairwise interactions), with several source nodes s_i, and sink nodes t_j, a reference regulatory network can be curated from literature or obtained from a public database. The steady state knock-down version of this problem has been shown to be NP-complete (Hashemikhabir et al., 2012). When the same problem is formulated as a linear optimization problem, the solution of this optimization problem provides a network, satisfying the experimental observations with minimum number of changes (insertion or deletion) of the edges on the reference network. The raw time-series expression data is assumed to be processed, and the binary activity data is available for the proteins in the network. We have used the cutoffs ($p < 0.05$ and variance <15) as described in the Datasets. Steiner nodes are also assumed to be active based on their presence in the reconstructed PCSF networks.

As the objective function of the model is to minimize the edit operations, i.e., insertions/deletions of edges, on the reference network, the proposed model also works when there is no reference network available. For such cases, the smallest network satisfying the expression data is sought. Let x_{ij} be the binary variable representing the presence of the edge from node i to node j in the reference network. If the edge is present, then the value of x_{ij} is 1, otherwise it is 0. Correspondingly, w_{ij} represents the presence of the edge from node i to node j in the network to be reconstructed from observations. For a graph G(V, E) with n nodes, the objective function is given in Equation (2), which basically quantifies the difference between the reference network and the reconstructed network.

In the solution phase, the matrix of state variables is used for the construction of the linear constraints. A protein is assumed to be activated once the corresponding state variable becomes 1. For the model, it does not matter what value the state variable is assigned thereafter. For the construction of the constraints, the kinematics of the system is taken into consideration. A protein is assumed to be activated by any protein, which is already activated at any previous time point and also a protein is able to activate any protein at any of the following time points. The constraints are based on the following assumptions: (1) sources are the proteins which are activated at the first time point and sinks are the proteins which are activated at the last time point, (2) each source node and sink node has to be connected to the network,

(3) at each time point, the proteins may only be activated by the upstream proteins, which are active in preceding time points, (4) no direct edges from sources to sinks are allowed, (5) no edges between sources or sinks are allowed, and (6) no self-edges are allowed. Note that these assumptions do not allow an upstream edge. However, there may be such edges in the reference network which are to be removed in the reconstructed network. Based on these assumptions, the following graphical constraints are derived.

1. There should be at least one edge going out of each source protein to the proteins activated at the second time point.
2. There should be at least one edge going into each sink proteins from the proteins activated at the last time point.
3. There should be at least one edge going into an intermediate node from the upstream nodes activated at a previous time point.
4. There should be at least one edge going out of an intermediate node to one of the downstream nodes, including the sink nodes.

Note that these constraints are derived only from the time series expression data. It is also possible to add additional constraints, if any perturbation experiment is available for the network. Let the set V_i be the set of proteins active at time point *i*. Let V_s be the set of source nodes and V_t be the set of target (i.e., sink) nodes. The node set of the reconstructed network V is the union of all source nodes, target nodes, and all the nodes active at some time point. Let V_p be the set of nodes activated just before the sink nodes. Let V_d be the set of downstream nodes that are activated after the activation of node *i*. The overall Integer Liner Program is then given as:

$$Minimize \sum_{i=1}^{n} \sum_{j=1}^{n} |x_{ij} - w_{ij}| \quad (2)$$

Subject to:

$$\sum\nolimits_{j \in V_1} x_{ij} \geq 1 \quad \text{for all } i \in V_s \quad (3)$$

$$\sum\nolimits_{i \in V_p} x_{ij} \geq 1 \quad \text{for all } j \in V_t \quad (4)$$

$$\sum\nolimits_{j \in V_d} x_{ij} \geq 1 \quad \text{for all } i \in V_x \quad (5)$$

$$\sum\nolimits_{i \in (V \setminus V_s)} x_{ij} \geq 1 \quad \text{for all } j \in V_{s+1} \quad (6)$$

Assessment of the Improvement

In the first step, we only use the temporal phosphoproteomic data and reconstruct the signaling network without a reference network. Then, PCSFs for each time point are merged together and a binary matrix has been created from the PCSF network in order to validate the edges and to determine edge directions the ILP based edge inference approach is used. The human protein interactome described in Datasets is assigned as the reference network for the ILP analysis. So, the PCSF and ILP based edge inference analysis are combined to provide the intermediate nodes (from the human interactome) based on the proteins identified at different time points from the experimental data, in addition to the direction information for the edges. The resulting directed network is then used for visualization and further analyses.

Network Analysis and Clustering

Restricted Neighborhood Search Cluster Algorithm (RNSC) in the NeAT toolbox (Brohee et al., 2008) was used to cluster the network where the maximum number of clusters was selected to be 20 and other parameters were kept as the default values. Critical nodes in the network were ranked by calculating four attributes: the path frequency, in-degree, out-degree, the sum of in-degree and out-degree, and the betweenness centralities. A simple path is an ordered sequence of nodes in a graph such that each node occurs at most once in this sequence and each pair of consecutive nodes is connected by an edge. Given all the possible simple paths between the terminal nodes at 2 min and the terminal nodes at 20 min, the path frequency of a node *p* is defined as the ratio of the simple paths that include *p* over all paths. The network analysis has been performed by using the Python NetworkX package (Schult and Swart, 2008).

Functional Enrichment Analysis

After clustering the network with RNSC algorithm, enrichment of each cluster has been assessed with DAVID web server in the following categories: biological process ontology, cellular component ontology, molecular function ontology, BBID pathways, BIOCARTA pathways, and KEGG pathways. Then, we have collected the enrichment results for each cluster and generated a matrix where rows are Gene Ontology (GO) terms and columns are corresponding *p*-values for each cluster, for all the data with $p < 0.05$, and enriched proteins >1%.

Results

Reconstruction of Temporal Signaling Networks in *Salmonella*-infected Human Cells

Modeling signaling networks is a more challenging task when the time dimension is taken into account. In a given temporal omic data, one of the problems to be solved is how the omic hits are connected and what are the upstream and downstream regulators in the final signaling network. The ILP based edge inference approach uses the temporal information as constraints and reconstructs edges and their direction between the omic hits toward solving this problem. But ILP-based edge inference approach cannot identify the missing components in the omic hits for the representation of the whole signaling system. The PCSF approach solves this problem by searching for the most confident region of the interactome that will include most of the experimental hits and adds hidden components, called Steiner nodes. As a limitation, if the initial interactome is undirected, the PCSF approach cannot assign directions to the edges and cannot use the temporal constraints, which can be solved by the ILP-based edge inference approach. As the different aspects of the ILP-based edge inference and PCSF network modeling approaches are complementing each other, we have combined these two approaches to reconstruct the temporal signaling networks in *Salmonella*-infected human cells.

In this work, we have used the phosphoproteomic data published in (Rogers et al., 2011). This dataset has been first divided into three parts at each time point based on the cellular compartment (membrane, cytosol, and nucleus). Additionally, the overlaps of the significantly phosphorylated proteins across different time points and different cellular compartments are compared. We have noted that the overlap between compartments within the same time points were very small, so the phosphoproteins at the same time point but in different compartments are safely merged (see Table S1). Next, for each time point, we have prepared a set of significantly phosphorylated proteins along with their fold changes during phosphorylation. To show the importance of revealing hidden components that were not present in the phosphoproteomic hit set, we first ran only the ILP-based edge inference approach on the data. The result was a disconnected network with many small sub-networks composed of three or four nodes which were not a representation of any pathway (Figure S1). This visual representation clearly showed that experimental hits alone are not enough to represent the complete network as there are missing components between these nodes. At this point, we took advantage of the PCSF approach in revealing hidden nodes and prepared a reference network to be used in the ILP-based edge inference approach. For this purpose, multiple PCSFs have been constructed for each time point and merged together to form a single network. To pipe this output into the ILP-based edge inference approach, we have prepared a network matrix where rows are proteins, columns are time points. When a node is present at a time point either as a phosphoproteomic hit or as a hidden node revealed by the PCSF approach, it is labeled as 1, otherwise it is 0. The ILP-based edge inference approach reconstructed a network based on the provided matrix as a reference network. The final network was composed of 658 nodes and 869 edges. As shown in **Figure 2**, the resulting network keeps the temporal information and also reveals hidden proteins and directions of the interactions. An interactive visualization and related source information is available at http://mistral.ii.metu.edu.tr/salmonella/salmonella_main.html. 547 out of 869 interactions were present in the reference human interactome. Remaining 322 interactions were novel, predicted interactions.

To check if this reconstructed network is specific to *Salmonella* infection, we have searched for the known targets of *Salmonella* effectors in the network. In this step, we have used the interactome that has been curated and compiled from published studies where 40 effectors interact with 50 human proteins through 62 interactions (Schleker et al., 2012). We found that 13 proteins out of 50 were present in the reconstructed network, which are known to be the targets of *Salmonella* effectors. The enrichment of *Salmonella* effector targets in the reconstructed network is statistically significant when compared to the overall human interactome ($p = 8.206 \times 10^{-8}$, by hypergeometric test) which implies the specificity of the reconstructed network to *Salmonella* infection. In **Table 1**, targets of *Salmonella* effectors present in the reconstructed network are listed with their functions and whether they are phoshoproteomic hits or found by our approach as intermediates. Additionally, we have checked the Gene Ontology and KEGG pathway enrichments for the overall reconstructed network. Regulation of transcription ($p = 2.0 \times 10^{-3}$), apoptosis ($p = 3.9 \times 10^{-10}$), intracellular transport ($p = 1.0 \times 10^{-8}$), cell cycle ($p = 2.1 \times 10^{-10}$), cytoskeleton organization ($p = 2.3 \times 10^{-17}$) are some of the processes enriched in the reconstructed network. More specifically, SNARE interaction in vesicular transport ($p = 1.0 \times 10^{-6}$), mTOR signaling ($p = 1.5 \times 10^{-4}$), and MAPK signaling ($p = 9.9 \times 10^{-6}$) pathways are among the enriched pathways. In several studies, *Salmonella* infection was shown to down-regulate mTOR pathway to induce apoptosis (Lee et al., 2014). The *Salmonella* effector AvrA targets MAPK signaling, mTOR signaling, and NF-κB pathways to manipulate the processes in the host cell (Liu et al., 2010).

Next, the reconstructed network is divided into 20 clusters using the RNSC algorithm, which searches highly connected node sets within a given graph. Each cluster was found to be enriched in few specific biological processes or pathways. Top three most significant processes are listed in **Table 2**. For example, the first cluster is enriched in the mTOR signaling pathway, cytoskeleton organization, and other processes. Also enrichments in the intracellular transport, apoptosis, RNA processing, and transcription is observed in different clusters. These clusters have also been analyzed based on the enrichment of cellular components, and number of clusters observed in three cellular compartments; nucleus, cytosol and cytoskeleton is reported in Figure S2.

With the help of network analysis techniques, we were able to rank the nodes present in the reconstructed network. The most central nodes (based on different measures) are listed in **Table 3** where 10 proteins that have known functions in apoptosis are observed. 14-3-3ζ (YWHAZ) and p53 (TP53) are the most frequently observed proteins on the simple paths passing through the hits from 2 min to 20 min. MAPK1, CDC42, MAP3K3, and 14-3-3δ (YWHAG) behave like signal transducers where the number of incoming edges are very low compared to other nodes; while MAPK1 also behaves like a signal receiver when the number of edges of each node were compared. These proteins are critical for the structure of the network; in other words, these proteins have the potential to reveal clinically important targets.

CDC42 is a Clinically Important Target in *Salmonella* Infection

Two effectors of *Salmonella*, SopE, and SopB, stimulate CDC42 (cell division control protein 42) protein which induces rearrangements in the cytoskeleton. CDC42 were not present in the set of phosphoproteomic hits; however, it was located in a central part of the reconstructed network, which is revealed by the PCSF approach. Although CDC42 has a phosphorylation site at tyrosine 64 (Tu et al., 2003) which is not present in the initial phosphoproteomic data, our approach correctly locates CDC42 in the final reconstructed network. Based on the total degrees, CDC42 is among the top 10 ranking proteins. In the reconstructed network, CDC42 has only one incoming edge, but has 13 outgoing edges which is consistent with the infection mechanism of *Salmonella* where stimulation of CDC42 leads to activation of many downstream signaling

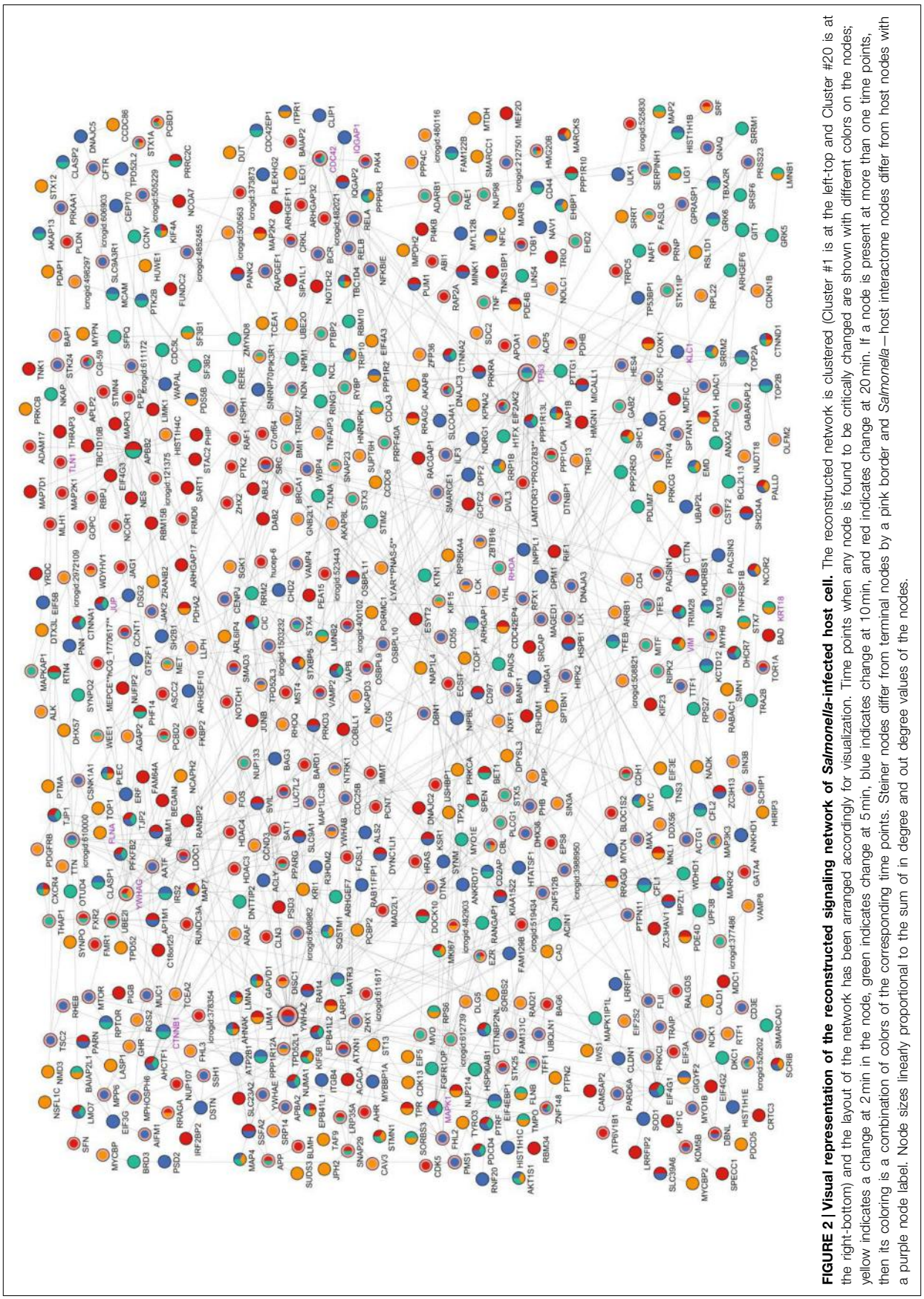


FIGURE 2 | Visual representation of the reconstructed signaling network of *Salmonella*-infected host cell. The reconstructed network is clustered (Cluster #1 is at the left-top and Cluster #20 is at the right-bottom) and the layout of the network has been arranged accordingly for visualization. Time points when any node is found to be critically changed are shown with different colors on the nodes; yellow indicates a change at 2 min in the node, green indicates change at 5 min, blue indicates change at 10 min, and red indicates change at 20 min. If a node is present at more than one time points, then its coloring is a combination of colors of the corresponding time points. Steiner nodes differ from terminal nodes by a pink border and *Salmonella*—host interactome nodes differ from host nodes with a purple node label. Node sizes linearly proportional to the sum of in degree and out degree values of the nodes.

TABLE 1 | Targets of *Salmonella* effectors in the reconstructed network.

Host protein	Function	Pathogen effectors	Type of the node in the network*
CDC42	Actin filament bundle assembly	SopB, SopE	intermediate
MTOR	Protein serine/threonine kinase activity	AvrA	intermediate
RHOA	Actin cytoskeleton organization	SifA, SseJ	intermediate
YWHAG	Negative regulation of protein serine/threonine kinase activity	SspH2	intermediate
TP53	Apoptotic activity	AvrA	intermediate
TLN1	Structural constituent of cytoskeleton	SseL	intermediate
KLC1	Microtubule motor activity	PipB2	pp-hit
FLNA	Actin cytoskeleton reorganization	SseI, SrfH, SspH2	pp-hit
CTNNB1	Cytoskeletal anchoring at plasma membrane	AvrA	pp-hit
VIM	Intermediate filament organization	SptP	pp-hit
IQGAP1	Ras GTPase activator activity	SseI, SrfH	pp-hit
JUP	Cytoskeletal anchoring at plasma membrane	SseF	pp-hit
KRT18	Intermediate filament cytoskeleton organization	SipC or SspC	pp-hit
MAPK1	Activation of MAPK activity	AvrA, SpvC	pp-hit
OSBPL9, OSBPL10, OSBPL11	Lipid transport	SseL	pp-hit (except OSBPL11)
STX1A, STX3, STX4, STX5, STX7, STX12, STXBP5	Intracellular protein transport		intermediate (except STX7, STX12)

**Intermediate, Steiner node; pp-hit, Phosphoproteomic hit.*

components including p21-activated kinases (PAKs) and PBD domain containing proteins (Galan and Zhou, 2000). PAK4 and PBD domain containing protein CDC42EP1 are among the downstream partners in the reconstructed network. Also, when we zoomed into the first degree neighbors of CDC42, their downstream components in our network can be observed (see **Figure 3A**). Some of these partners are active at 5 min, or at 10 min, or at other time points. The CDC42 shows a hub-like character in the reconstructed network. Hub proteins cannot interact with all their partners at the same time. They either adapt multiple binding sites or use a single binding site repeatedly. This property of hubs has been well-established for TP53 protein where four binding sites are repeatedly used to interact with different partners (Tuncbag et al., 2009). We have checked this property in CDC42 to understand its interactions by searching for the available structural data in Protein Databank (PDB) (Berman et al., 2000) and Interactome3D (Mosca et al., 2013). We have found six interactions out of 13 in atomic detail (see **Figure 3B**). Structural data provides information about at which region two proteins are interacting. Analysis of the binding site for each protein pair has shown that CDC42 is using the same binding region completely or partially to interact with its partners and this property is a characteristic of hub proteins. Also, the downstream partners of CDC42 are active at different time points as illustrated in **Figure 3A**, which also shows the mutually exclusive character of the interactions. For example, PAK4 is in the reconstructed network showing its effect after infection at time points 10 and 20 min. PARD6A and ARHGAP32 effect at 10 and 20 min, respectively. The partner proteins are effective at different time points and CDC42 can bind to these proteins in a mutually exclusive manner. IQGAP1 is another downstream component, which promotes *Salmonella* invasion by binding to CDC42 and knockdown of IQGAP1 was shown to be reducing the invasion (Brown et al., 2007).

The Reconstructed Signaling Network Revealed Many Other Potential Clinical Targets

Besides CDC42, some other targets of *Salmonella* effectors were located correctly in the reconstructed network although they were not present in the initial phosphoproteomic data; such as 14-3-3δ, RHOA, TP53, TLN1 (Schleker et al., 2012). Also pathogen targets such as β-catenin, MAPK1, IQGAP1 are observed in the reconstructed network as phosphoproteomic hits (Schleker et al., 2012).

In addition to these targets, mTOR pathway was found to be enriched in the reconstructed network. mTOR signaling was known to be altered after *Salmonella* infection and mTOR is a phosphoproteomic hit having significant effect at 10 min in our data. When we have investigated the neighbors of the mTOR protein (**Figure 4A**) RHEB, a direct regulator of mTOR (Long et al., 2005), is observed as an interactor of mTOR in the reconstructed network. Also, RPTOR binding to mTOR and EIF4EBP1 was recovered in our network. The mTOR - RHEB complex induces phosphorylation of EIF4EBP1 (Long et al., 2005). Even though RHEB is an important player in the activation of EIF4EBP1, it was not observed within the initial phosphoproteomic hits. The proposed two-step modeling approach was able to locate RHEB in the final network, completing the missing interactions of the signaling pathway. According to our network, signaling in the RPTOR-mTOR-Rheb-EIF4EBP1 axis starts at 5 min and continues until 10 min.

RHOA, is a GTPase that functions in the actin cytoskeleton organization (Hall, 1998). It was a Steiner node in final network as a target of *Salmonella* effectors. It has 5 binding partners in the reconstructed network of where four of them are

TABLE 2 | Gene ontology (GO) biological process enrichments of each cluster located in the final network.

Cluster #	GO Term	*p*-values	Percent
1	Cytoskeleton organization	0.00180	1.65
	Regulation of cellular component size	0.00220	1.37
	Regulation of cytoskeleton organization	0.00270	1.1
2	Cytoskeleton organization	0.00906	1.5
	Regulation of phosphorylation	0.01138	1.5
3	Protein amino acid phosphorylation	0.01509	1.45
	Phosphorylation	0.02753	1.45
4	RNA processing	0.0013	1.94
	Intracellular signaling cascade	0.00152	2.77
	Cell cycle process	0.00155	1.94
5	Not any significant GO enrichment		
6	Intracellular transport	0.00233	1.65
	Membrane organization	0.00842	1.18
	Negative regulation of macromolecule metabolic process	0.0186	1.41
7	Response to organic cyclic substance	0.001745	1.03
	Negative regulation of macromolecule metabolic process	0.002774	1.81
	Regulation of cell cycle	0.00395	1.29
8	Positive regulation of specific transcription from RNA polymerase II promoter	0.00381	1.05
	Cell death	0.005171	2.11
	Death	0.005325	2.11
9	RNA processing	0.002175	1.72
	Transmembrane receptor protein tyrosine kinase signaling pathway	0.0022984	1.23
	Cytoskeleton organization	0.004371	1.47
10	Regulation of small GTPase mediated signal transduction	3.3120E-9	1.13
	Small GTPase mediated signal transduction	3.52554E-7	1.01
	Intracellular signaling cascade	5.2661E-4	1.13
11	Regulation of cellular protein metabolic process	0.00220	1.57
	Response to organic substance	0.002536	1.83
	Response to hormone stimulus	0.005706	1.31
12	Actin filament-based process	0.008572	1.1
	Cell cycle phase	0.035767	1.1
	Cytoskeleton organization	0.04074	1.1
13	Negative regulation of programmed cell Death	0.00105	1.73
	Negative regulation of cell death	0.00106	1.73
	Negative regulation of cell proliferation	0.007898	1.45
14	M phase	0.003363	1.12
	Cell cycle phase	0.007577	1.12
	Chromosome organization	0.013039	1.12
15	Protein import	0.001547	1.21
	Protein localization in organelle	0.0021107	1.21
	Intracellular transport	0.005237	1.82

(Continued)

TABLE 2 | Continued

Cluster #	GO Term	*p*-values	Percent
16	Macromolecular complex assembly	0.012584	1.04
	Macromolecular complex subunit organization	0.0163	1.04
	Regulation of apoptosis	0.02648	1.04
17	Tube development	0.005909	1.38
	Tube morphogenesis	0.01944	1.04
	Cellular component morphogenesis	0.0287436	1.38
18	Regulation of gene-specific transcription	0.00186	1.29
	Positive regulation of macromolecule metabolic process	0.00397	2.26
	Positive regulation of macromolecule biosynthetic process	0.00617	1.94
19	Positive regulation of molecular function	0.016011	1.69
	Positive regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process	0.019733	1.69
	Positive regulation of nitrogen compound metabolic process	0.021893988499693085	1.69
20	Not any significant GO enrichment		

The percent column has been calculated by dividing the number of involved proteins in that GO term to the total number of proteins in that cluster and converted into percentage.

upstream interactors and only one is a downstream interactor (**Figure 4B**). Among the upstream interactors, RPS6KA4 is effective at 2 min and the remaining ones are active at 5 min. The downstream interactor INPPL1 is effective at 10 min. So, the final reconstructed network suggests that RHOA receives signals from proteins active at min 2 and min 5 and transmits these signals until min 10.

The 14-3-3δ protein (YWHAG) shows a pattern similar to CDC42 where the incoming interactions are not present, but there are many outgoing interactions from 14-3-3δ which implies that 14-3-3δ is a signal mediator for the downstream components of the network. In **Figure 4C**, first neighbors of 14-3-3δ are illustrated in the network. 14-3-3δ is a Steiner node, a known target of *Salmonella* effectors, and it is effective at min 2, 5, and 20. Its 13 outgoing edges suggest a function like a signal transducer, sending signals to many downstream proteins at different time points.

Also, seven proteins from the Syntaxin family were present in the reconstructed network and only three of them was a phosphoproteomic hit, others were Steiner nodes found by our approach. Syntaxins function in vesicle trafficking which is an important process in *Salmonella* replication and transport. *Salmonella* effectors hijack syntaxins by binding them (Ramos-Morales, 2012).

Finally, oxysterol-binding proteins (OSBPs) are also present in the reconstructed network, which are known to be enhancing replication of *Salmonella* in the host cell by interacting with the *Salmonella* effector SseL (Auweter et al., 2012). This interaction can lead to the exploitation of OSBP dependent pathways altered during the *Salmonella* infection.

Discussion

Improvements in the high-throughput technologies revolutionized the systems biology era. Instead of comparing lists of genes or proteins, finding the interactions, regulations, and mechanisms within a set of significantly altered proteins or genes have gained importance. In addition, integration of multiple "omic" hits in a biologically meaningful way is now crucial to better understand the functional pathways and cellular mechanisms that are active during a disease or perturbation. For this purpose, several network modeling approaches have been developed, which successfully reveals clinically valuable targets and important pathways, especially in several cancer types. Another dimension of omic data is its temporality, which makes the network modeling process more challenging, as instead of simply connecting omic hits, the time related constraints have to be considered during network reconstruction.

In this work, we have provided a proof-of-concept application of an integrative approach which benefits from two different network reconstruction methods, namely PCSF and ILP based edge inference methods. Although both methods infer networks from experimental omic hits, they perform better in different parts of the modeling. The former reveals the hidden components of the signaling, but cannot handle time as a dimension. The latter can integrate temporal information to reconstruct directed edges, but cannot add missing signaling components to complete the lacking parts of the signaling. These complementary aspects of the methods inspired us to combine both approaches to model the signaling network of human cells after *Salmonella* infection based on the temporal phosphoproteomic data. We have selected *Salmonella* infection, because the signaling changes in the host cells are still unknown despite the efforts to understand the communication details between the pathogen and the host. In addition, available approaches have not been yet applied to model signaling changes in the host organism during *Salmonella* infection. In the first stage, we have integrated the temporal phosphoprotemic data of *Salmonella*-infected human cells with confidence weighted protein–protein interactions to reconstruct the signaling pathway for each time point with

TABLE 3 | Top ranking proteins in the reconstructed network.

Name	Function	Subcellular location	Path frequency	In degree	Out degree	Betweenness
ACTG1	Structural constituent of cytoskeleton	Actin cytoskeleton	7.3	2	5	0.0017
AIFM1	Apoptotic process	Mitochondrion	6.8	1	1	0.0017
APBB2	Beta-amyloid binding	Cytoplasm	0.0	1	24	0.0001
ATXN1	DNA binding	Cytoplasm, nucleus	14.8	8	5	0.0021
CDC42	Actin cytoskeleton organization	Cytoskeleton	4.2	1	13	0.0002
CFL1	Actin cytoskeleton organization	Cytoskeleton	24.3	8	1	0.001
CIC	DNA binding	Nucleus	9.5	5	1	0.0008
CRTC3	cAMP response element binding protein binding	Cytoplasm, nucleus	12.7	1	0	0.0
CTNNB1	Alpha-catenin binding	Cytoskeleton	0.0	2	9	0.0008
EIF2AK2	Protein serine/threonine kinase activity	Cytoplasm, nucleus	1.6	6	2	0.001
EIF3G	Translation initiation factor activity	Cytoplasm, nucleus	6.8	2	2	0.002
EIF4G1	Translation factor activity, nucleic acid binding	Cytosol, membrane	37.6	8	6	0.004
HSPB1	Cellular component movement	Cytoskeleton	22.0	3	2	0.0012
MAP3K3	Activation of MAPKK activity	Cytosol	0.0	0	9	0.0
MAPK1	MAP kinase activity	Cytoskeleton	11.1	12	1	0.0015
MPP6	Maturation of 5.8S rRNA	Membrane	9.8	2	2	0.002
MPZL1	Cell-cell signaling	Membrane	24.9	3	1	0.0005
RELA	Sequence-specific DNA binding transcription factor activity	Nucleus	0.0	3	11	0.0016
SRC	Protein tyrosine kinase activity	Membrane, cytoskeleton	2.1	3	13	0.0006
TOP2A	Chromatin binding	Cytoplasm, nucleus	0.0	6	1	0.0006
TP53	Tumor suppressor; induces growth arrest or apoptosis	Cytoplasm, nucleus	32.8	15	14	0.011
UBE2I	SUMO ligase activity	Cytoplasm, nucleus	7.3	4	8	0.0016
YWHAG	Protein kinase binding	Cytoplasmic vesicle membrane	0.0	0	13	0.0
YWHAZ	Protein kinase binding	Cytoplasmic vesicle membrane	65.1	34	11	0.0093
ZC3HAV1	Cellular response to exogenous dsRNA	Cytoplasm, nucleus	24.9	1	0	0.0

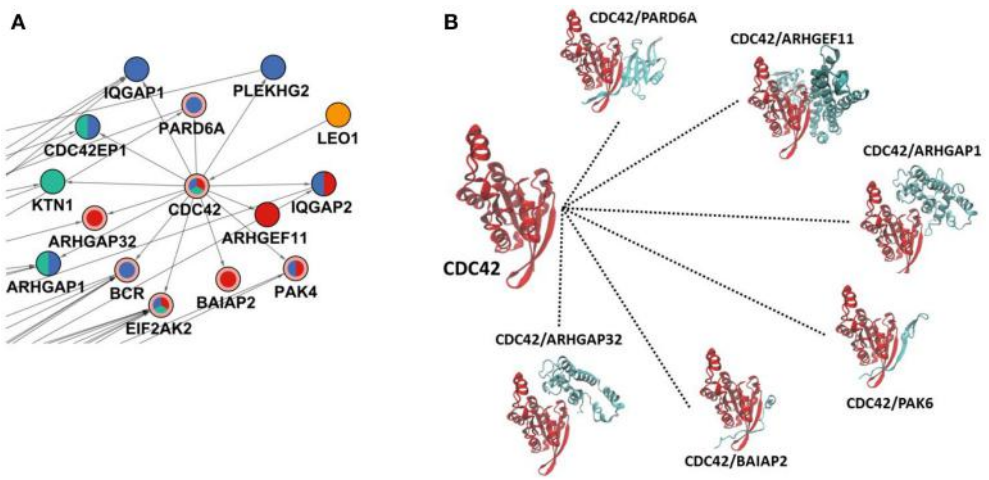


FIGURE 3 | CDC42 and its interactions in the reconstructed network. (A) The region where CDC42 and its first neighbors are located in the reconstructed network. The coloring scheme is the same as in **Figure 2** where CDC42, EIF2AK2, BCR, BAIAP2, ARHGAP32, PARD6A, and PAK4 are Steiner nodes and others are phosphoproteomic hits. **(B)** Structural details of CDC42 interactions where CDC42 uses the same binding site completely or partially to interact with its partners. Here, PAK6 is a structural homolog of PAK4; therefore to show similar binding of PAK4, it is represented with PAK6.

the PCSF approach. Then, all the components were labeled with the corresponding time points based on their presence in the reconstructed networks, and used as the input for the ILP-based edge inference step, in which directions are assigned to the interactions based on the temporal data. The final network with 658 proteins and 869 interactions provided a rich

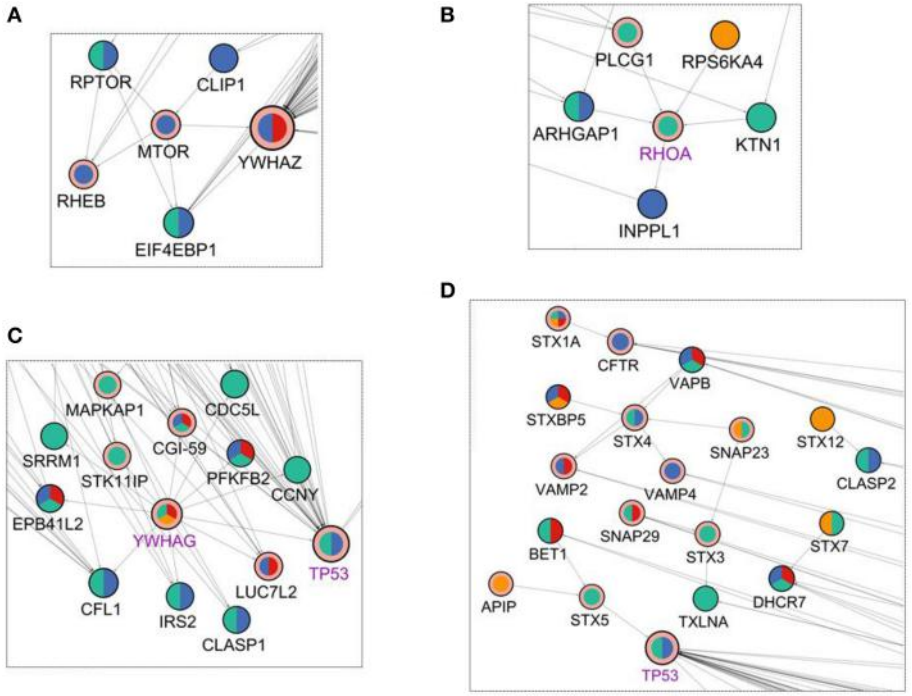


FIGURE 4 | Visualization of the first degree neighbors of (A) mTOR, (B) RHOA, (C) YWHAG, and (D) Syntaxins in the reconstructed network in Figure 2. The coloring scheme is the same as in **Figure 2**.

source to analyze the signaling alterations and clinical target identification. Our approach allowed us to identify host pathways functioning during the *Salmonella* infection and to rank the proteins according to their importance for the infection based on their centrality in the network. The resulting network conserves the information about temporality, direction of interactions, while revealing the hidden entities in the signaling. Several pathways such as SNARE binding, mTOR signaling, immune response, cytoskeleton organization, and apoptosis, were found to be effected, many of which were previously found to be altered in the host cell after *Salmonella* infection. Additionally, we have shown that the reconstructed network is enriched in the protein targets of the *Salmonella* effectors. Clustering of the resulting network showed that the multiple biological processes are enriched in each cluster. The final network also involves enrichments in the cytoskeletal organization and the regulation of cellular component size. These findings are in parallel with the known infection mechanism of the *Salmonella* where the injected effector proteins trigger the epithelial cell membrane by rearranging the cytoskeleton of the host cell that results in invasion of the bacterium into the host cell.

Another benefit of the proposed two-step approach (**Figure 1**) was that hidden components of signaling can be revealed with network reconstruction. In this specific demonstration, several known targets of *Salmonella* effectors have been accurately included in the reconstructed network such as CDC42, RHOA, 14-3-3δ, Syntaxins although they were not present in the initial phosphoproteomic data. These hidden signaling components can be potential therapeutic targets. Among them CDC42 is a target of the effector protein SopB and their interaction helps in the adaptation of *Salmonella* to the intracellular condition of the host. CDC42 is responsible of downstream signaling and behaves as a signal transmitter. From a medical point of view, targeting CDC42 is a good approach both for blocking the adaptation of *Salmonella* in the host cell and abnormal downstream signaling during infection (**Figure 3**). RHOA functions in cytoskeleton organization and also it a target of *Salmonella* effectors. The effector SifA activates RHOA during infection. Activated RHOA promotes opening tubes in the membrane (Srikanth et al., 2010). Therefore, RHOA can be considered as a therapeutic target and controlling its activation can be a good approach in *Salmonella* treatment (**Figure 4B**). Also, besides revealing the hidden components, the reconstructed edge directions nicely showed how the signals are transmitted temporally from one layer to another. For example, some host proteins behave like a signal receiver such as MAPK1 and some others behave like a signal transmitter such as, 14-3-3δ (**Figure 4C**).

Understanding these communications and signaling details in the host is crucial to improve the available treatment strategies for *Salmonella* infection in the near future, especially as the new antibiotic-resistance species are on the rise. We believe that the integrated approaches, such as the one presented here, have a high potential for understanding the key molecular mechanisms in bacteria's susceptibility or resistance to the available antibiotics and for the identification of new clinical

targets in infectious diseases, especially in the *Salmonella* infection.

Acknowledgments

NT thanks the TUBITAK-Marie Curie Co-funded Brain Circulation Scheme (project no. 114C026) for the support. OEO is partially supported by the Scientific and Technological Research Council of Turkey (TÜBITAK) 1002 Short Term Research and Development Funding Program Grant #113E323.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00730

References

Aijo, T., Granberg, K., and Lahdesmaki, H. (2013). Sorad: a systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements. *Bioinformatics* 29, 1283–1291. doi: 10.1093/bioinformatics/btt130

Auweter, S. D., Yu, H. B., Arena, E. T., Guttman, J. A., and Finlay, B. B. (2012). Oxysterol-binding protein (OSBP) enhances replication of intracellular Salmonella and binds the Salmonella SPI-2 effector SseL via its N-terminus. *Microbes Infect.* 14, 148–154. doi: 10.1016/j.micinf.2011.09.003

Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., Francois, J. M., et al. (2010). Finding undetected protein associations in cell signaling by belief propagation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 882–887. doi: 10.1073/pnas.1004751108

Bender, C., Henjes, F., Frohlich, H., Wiemann, S., Korf, U., and Beissbarth, T. (2010). Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics* 26, i596–i602. doi: 10.1093/bioinformatics/btq385

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., et al. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.* 36, W444–W451. doi: 10.1093/nar/gkn336

Brown, M. D., Bry, L., Li, Z., and Sacks, D. B. (2007). IQGAP1 regulates Salmonella invasion through interactions with actin, Rac1, and Cdc42. *J. Biol. Chem.* 282, 30265–30272. doi: 10.1074/jbc.M702537200

Dandekar, T., Astrid, F., Jasmin, P., and Hensel, M. (2012). *Salmonella enterica*: a surprisingly well-adapted intracellular lifestyle. *Front. Microbiol.* 3:164. doi: 10.3389/fmicb.2012.00164

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9, 67–103. doi: 10.1089/10665270252833208

Dhal, P. K., Barman, R. K., Saha, S., and Das, S. (2014). Dynamic modularity of host protein interaction networks in Salmonella Typhi infection. *PLoS ONE* 9:e104911. doi: 10.1371/journal.pone.0104911

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231. doi: 10.1093/bioinformatics/btn161

Eckmann, L., Smith, J. R., Housley, M. P., Dwinell, M. B., and Kagnoff, M. F. (2000). Analysis by high density cDNA arrays of altered gene expression in human intestinal epithelial cells in response to infection with the invasive enteric bacteria Salmonella. *J. Biol. Chem.* 275, 14084–14094. doi: 10.1074/jbc.275.19.14084

Eren Ozsoy, O., Aydin Son, Y., and Can, T. (2015). *Reconstruction of Signaling and Regulatory Networks using Time-series and Perturbation Experiments*. Technical Report, Department of Computer Engineering, Middle East Technical University.

Eren Ozsoy, O., and Can, T. (2013). A divide and conquer approach for construction of large-scale signaling networks from PPI and RNAi data using linear programming. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 10, 869–883. doi: 10.1109/TCBB.2013.80

Frohlich, H., Sahin, O., Arlt, D., Bender, C., and Beissbarth, T. (2009). Deterministic effects propagation networks for reconstructing protein signaling networks from multiple interventions. *BMC Bioinformatics* 10:322. doi: 10.1186/1471-2105-10-322

Galan, J. E., and Zhou, D. (2000). Striking a balance: modulation of the actin cytoskeleton by Salmonella. *Proc. Natl. Acad. Sci. U.S.A.* 97, 8754–8761. doi: 10.1073/pnas.97.16.8754

Gosline, S. J., Spencer, S. J., Ursu, O., and Fraenkel, E. (2012). SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. *Integr. Biol. (Camb).* 4, 1415–1427. doi: 10.1039/c2ib20072d

Hall, A. (1998). Rho GTPases and the actin cytoskeleton. *Science* 279, 509–514. doi: 10.1126/science.279.5350.509

Hashemikhabir, S., Ayaz, E. S., Kavurucu, Y., Can, T., and Kahveci, T. (2012). Large-scale signaling network reconstruction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9, 1696–1708. doi: 10.1109/TCBB.2012.128

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems* 96, 86–103. doi: 10.1016/j.biosystems.2008.12.004

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, S. S., Clarke, D. C., Gosline, S. J. C., Labadorf, A., Chouinard, C. R., Gordon, W., et al. (2012). Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. *PLoS Comp Biol.* 9:e1002887. doi: 10.1371/journal.pcbi.1002887

Huang, S. S., and Fraenkel, E. (2009). Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci.Signal.* 2, ra40. doi: 10.1126/scisignal.2000350

Kiani, N. A., and Kaderali, L. (2014). Dynamic probabilistic threshold networks to infer signaling pathways from time-course perturbation data. *BMC Bioinformatics* 15:250. doi: 10.1186/1471-2105-15-250

Kim, Y. A., Wuchty, S., and Przytycka, T. M. (2011). Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput. Biol.* 7:e1001095. doi: 10.1371/journal.pcbi.1001095

Knapp, B., and Kaderali, L. (2013). Reconstruction of cellular signal transduction networks using perturbation assays and linear programming. *PLoS ONE* 8:e69220. doi: 10.1371/journal.pone.0069220

Lawhon, S. D., Khare, S., Rossetti, C. A., Everts, R. E., Galindo, C. L., Luciano, S. A., et al. (2011). Role of SPI-1 secreted effectors in acute bovine response to *Salmonella enterica* Serovar Typhimurium: a systems biology analysis approach. *PLoS ONE* 6:e26869. doi: 10.1371/journal.pone.0026869

Lee, C. H., Lin, S. T., Liu, J. J., Chang, W. W., Hsieh, J. L., and Wang, W. K. (2014). Salmonella induce autophagy in melanoma by the downregulation of AKT/mTOR pathway. *Gene Ther.* 21, 309–316. doi: 10.1038/gt.2013.86

Linde, J., Schulze, S., Henkel, S., and Guthke, R. (2015). Data- and knowledge-based modeling of gene regulatory networks: an update. *EXCLI J.* 346–378. doi: 10.17179/excli2015-168

Liu, X., Lu, R., Xia, Y., Wu, S., and Sun, J. (2010). Eukaryotic signaling pathways targeted by Salmonella effector protein AvrA in intestinal infection *in vivo*. *BMC Microbiol.* 10:326. doi: 10.1186/1471-2180-10-326

Long, X., Lin, Y., Ortiz-Vega, S., Yonezawa, K., and Avruch, J. (2005). Rheb binds and regulates the mTOR kinase. *Curr. Biol.* 15, 702–713. doi: 10.1016/j.cub.2005.02.053

Markowetz, F., Kostka, D., Troyanskaya, O. G., and Spang, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* 23, i305–i312. doi: 10.1093/bioinformatics/btm178

Matos, M. R., Knapp, B., and Kaderali, L. (2015). lpNet: a linear programming approach to reconstruct signal transduction networks. *Bioinformatics*. doi: 10.1093/bioinformatics/btv327. [Epub ahead of print].

Melas, I. N., Samaga, R., Alexopoulos, L. G., and Klamt, S. (2013). Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput. Biol.* 9:e1003204. doi: 10.1371/journal.pcbi.1003204

Mosca, R., Ceol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. *Nat. Methods* 10, 47–53. doi: 10.1038/nmeth.2289

Raghunathan, A., Reed, J., Shin, S., Palsson, B., and Daefler, S. (2009). Constraint-based analysis of metabolic capacity of *Salmonella typhimurium* during host-pathogen interaction. *BMC Syst. Biol.* 3:38. doi: 10.1186/1752-0509-3-38

Ramos-Morales, F. (2012). Impact of *Salmonella enterica* type III secretion system effectors on the eukaryotic host cell. *ISRN Cell Biol.* 2012, 1–36. doi: 10.5402/2012/787934

Rogers, L. D., Brown, N. F., Fang, Y., Pelech, S., and Foster, L. J. (2011). Phosphoproteomic analysis of Salmonella-infected cells identifies key kinase regulators and SopB-dependent host phosphorylation events. *Sci. Signal.* 4, rs9. doi: 10.1126/scisignal.2001668

Schleker, S., Sun, J., Raghavan, B., Srnec, M., Müller, N., Koepfinger, M., et al. (2012). The current Salmonella-host interactome. *Proteomics. Clin. Appl.* 6, 117–133. doi: 10.1002/prca.201100083

Schult, D. A., and Swart, P. (2008). "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, eds G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA), 11–16.

Srikanth, C., Wall, D. M., Maldonado-Contreras, A., Shi, H. N., Zhou, D., Demma, Z., et al. (2010). Salmonella pathogenesis and processing of secreted effectors by caspase-3. *Science* 330, 390–393. doi: 10.1126/science.1194598

Steele-Mortimer, O. (2011). Exploitation of the ubiquitin system by invading bacteria. *Traffic* 12, 162–169. doi: 10.1111/j.1600-0854.2010.01137.x

Stolovitzky, G., Monroe, D., and Califano, A. (2007). Dialogue on Reverse—Engineering Assessment and Methods. *Ann. N. Y. Acad. Sci.* 1115, 1–22. doi: 10.1196/annals.1407.021

Taylor, R. C., Singhal, M., Weller, J., Khoshnevis, S., Shi, L., and Mcdermott, J. (2009). A network inference workflow applied to virulence-related processes in Salmonella typhimurium. *Ann. N. Y. Acad. Sci.* 1158, 143–158. doi: 10.1111/j.1749-6632.2008.03762.x

Tu, S., Wu, W. J., Wang, J., and Cerione, R. A. (2003). Epidermal growth factor-dependent regulation of Cdc42 is mediated by the Src tyrosine kinase. *J. Biol. Chem.* 278, 49293–49300. doi: 10.1074/jbc.M307021200

Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S. S., Chayes, J., Borgs, C., et al. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J. Comput. Biol.* 20, 124–136. doi: 10.1089/cmb.2012.0092

Tuncbag, N., Kar, G., Gursoy, A., Keskin, O., and Nussinov, R. (2009). Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example. *Mol. Biosyst.* 5, 1770–1778. doi: 10.1039/b905661k

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010:baq023. doi: 10.1093/database/baq023

Yeger-Lotem, E., Riva, L., Su, L. J., Gitler, A. D., Cashikar, A. G., King, O. D., et al. (2009). Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat. Genet.* 41, 316–323. doi: 10.1038/ng.337

Yoon, H., McDermott, J. E., Porwollik, S., McClelland, M., and Heffron, F. (2009). Coordinated regulation of virulence during systemic infection of *Salmonella enterica* serovar Typhimurium. *PLoS Pathog.* 5:e1000306. doi: 10.1371/journal.ppat.1000306

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in
MICROBIOLOGY

REVIEW ARTICLE
published: 24 February 2015
doi: 10.3389/fmicb.2015.00094

Computational approaches for prediction of pathogen-host protein-protein interactions

***Esmaeil Nourani*[1], *Farshad Khunjush*[1,2]* *and Saliha Durmuş*[3]**

[1] *Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran*
[2] *School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[3] *Computational Systems Biology Group, Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey*

Edited by:
Evangelos Giamarellos-Bourboulis, University of Athens, Medical School, Greece

Reviewed by:
Magdalena Chirila, Iuliu Hatieganu University of Medicine and Pharmacy, Romania
Malgorzata Anna Mikaszewska-Sokolewicz, The Medical University of Warsaw, Poland

***Correspondence:**
Farshad Khunjush, Department of Computer Science and Engineering, School of Electrical and Computer Engineering, Zand Avenue, Shiraz 71348 - 51154, Iran
e-mail: khunjush@shirazu.ac.ir

Infectious diseases are still among the major and prevalent health problems, mostly because of the drug resistance of novel variants of pathogens. Molecular interactions between pathogens and their hosts are the key parts of the infection mechanisms. Novel antimicrobial therapeutics to fight drug resistance is only possible in case of a thorough understanding of pathogen-host interaction (PHI) systems. Existing databases, which contain experimentally verified PHI data, suffer from scarcity of reported interactions due to the technically challenging and time consuming process of experiments. These have motivated many researchers to address the problem by proposing computational approaches for analysis and prediction of PHIs. The computational methods primarily utilize sequence information, protein structure and known interactions. Classic machine learning techniques are used when there are sufficient known interactions to be used as training data. On the opposite case, transfer and multitask learning methods are preferred. Here, we present an overview of these computational approaches for predicting PHI systems, discussing their weakness and abilities, with future directions.

Keywords: protein-protein interaction, pathogen-host interaction (PHI), computational PHI prediction, machine learning, data mining

INTRODUCTION

Many studies concerning identification of protein interactions and their associated networks were published (Aloy and Russell, 2003). Most of the previous studies were primarily focused on determining protein-protein interactions (PPIs) within a single organism (intra-species PPI prediction), while the prediction of PPIs between different organisms (inter-species PPI prediction) has recently emerged. Inter-species interactions may take many forms; in this survey, however, we focus on PPIs between pathogens and their hosts. Pathogen-host interaction (PHI) prediction is worthwhile to enlighten the infection mechanisms in the scarcity of experimentally-verified PHI data. Interactions between pathogen and host proteins allow pathogenic microorganisms to manipulate host mechanisms in order to use host capabilities and to escape from host immune responses (Dyer et al., 2010). Therefore, a complete understanding of infection mechanisms through PHIs is crucial for the development of new and more effective therapeutics.

Despite the critical need to improve the PHI knowledge, current progress is not adequate, suffering from scarcity of available experimental PHI data. Reliable experimental methods are time-consuming and expensive, making it unjustifiable to evaluate all possible PHIs. For instance, considering about 26,000 human proteins paired with a few thousands of pathogen proteins lead to millions of protein pairs to test experimentally. Scarce verified interactions are collected within a number of databases like HPIDB (Kumar and Nanduri, 2010), PATRIC (Wattam et al., 2014), PHISTO (Durmuş Tekir et al., 2013), VirHostNet (Navratil et al., 2009), and VirusMentha (Calderone et al., 2014). At this point, computational approaches come to help by predicting putative PHIs. In this paper, we concentrate on these computational studies, which are mandatory for enriching the available data and consequently increasing the pace of research in the field. The methods which were successfully applied specifically for PHI prediction in the literature are categorized based on pathogen-host systems in **Table 1**.

Considering the relative availability of interaction data for HIV-Human system, notable number of studies are dedicated to this pathogen. Some other viral and bacterial pathogens are investigated and human is the main target as the host for investigation. Computational methods for predicting PHIs exploit known protein and domain interactions, and information on sequence of proteins. Network topology measures can complement these data. For instance, targeting hubs and bottleneck proteins in human PPI network by pathogen proteins is a well-accepted idea (Dyer et al., 2008; Durmuş Tekir et al., 2012; Schleker and Trilling, 2013; Zheng et al., 2014), though, they are not the sole targeted proteins (Chen et al., 2012). Classic machine learning methods are valuable remedy for cases where enough data for training are available. However, valuable efforts have recently been performed to apply these techniques for situations suffer from scarcity of known interaction data using machine learning based methods as transfer and multitask learning (Xu et al., 2010; Kshirsagar et al., 2013a,b).

In PPI prediction studies, methods specific for intra-species interactions are usually used. On the other hand, concentrating

Table 1 | Computational studies for prediction of PHIs.

Pathogen-host system	References
Plasmodium falciparum-Human	Krishnadev and Srinivasan, 2008 Lee et al., 2008 Wuchty, 2011 Dyer et al., 2007
Helicobacter pylori-Human	Kim et al., 2007; Tyagi et al., 2009
Hepatitis C virus (HCV)-Human	Cui et al., 2012; Zheng et al., 2014
Phage T4-*Escherichia coli*	Krishnadev and Srinivasan, 2011
Phage lambda-*E. coli*	Krishnadev and Srinivasan, 2011
C. albicans-Zebrafish	Wang et al., 2013
E. coli-Human	Krishnadev and Srinivasan, 2011
Plasmodium berghei-Mouse	Reid and Berriman, 2013
Plasmodium berghei-Insect vector (Mosquito)	Reid and Berriman, 2013
Oral microbial-Human	Coelho et al., 2014
Salmonella-Human	Krishnadev and Srinivasan, 2011 Arnold et al., 2012 Kshirsagar et al., 2012 Kshirsagar et al., 2013b Schleker et al., 2012a Mei and Zhu, 2014 Schleker et al., 2014 (Review)
Mycobacterium Tuberculosis H37Rv-Human	Zhou et al., 2014
Yersinia pestis-Human	Krishnadev and Srinivasan, 2011 Kshirsagar et al., 2012 Kshirsagar et al., 2013b
Mycobacterium apicomplexa and *Mycobacterium kinetoplastida*-Human	Davis et al., 2007
Xanthomonas oryzae-Rice	Kim et al., 2008
HTLV-Human	Mei, 2014
HIV1-Human	Evans et al., 2009 Tastan et al., 2009 Mei, 2013 Qi et al., 2010 Dyer et al., 2011 Ray et al., 2012 Doolittle and Gomez, 2010
	Nouretdinov et al., 2012 Mukhopadhyay et al., 2010, 2012, 2014 Mondal et al., 2012

(Continued)

Table 1 | Continued

Pathogen-host system	References
36 viral species-Human	Franzosa and Xia, 2011
Influenza A NS1–Human	De Chassey et al., 2013
HPV16–Human	Dong et al., in press
Bacillus anthracis-Human	Kshirsagar et al., 2013b
Francisella tularensis-Human	Kshirsagar et al., 2013b
Dengue virus-Human	Doolittle and Gomez, 2011 Segura-Cabrera et al., 2013
Insect vector *A. aegypti*-Human	Doolittle and Gomez, 2011
Salmonella-Arabidopsis	Schleker et al., 2012a Schleker et al., 2014 (Review)
Human papilloma viruses (HPV)-Human	Cui et al., 2012
R. solanacearum-Arabidopsis	Li et al., 2012
Y. pestis, M. tuberculosis, C. diphtheriae, C. ulcerans, E. coli, and C. pseudotuberculosis-Human, Goat, Sheep, and Horse	Barh et al., 2013

on the interactions between different organisms is a young branch of this field. The traditional methods cannot be applied here, their adaptation or devising new approaches would be mandatory.

MACHINE LEARNING AND DATA MINING BASED APPROACHES

Applying machine learning techniques to bioinformatics is a well-accepted idea (Baldi and Brunak, 2001), which includes early efforts for PPI predictions (Bock and Gough, 2001). These methods utilize available PPI data as features for training and classifying interacting and non-interacting protein pairs. Both semi-supervised and supervised learning are used for PHI prediction. A Supervised method, which exploits exclusively labeled data, is applied in Tastan et al. (2009) integrating 35 features within eight groups using Random Forest (RF) classifier to deal with noisy and redundant features. The semi-supervised extension of their work is presented in Qi et al. (2010) which discarded 17 attributes from the feature vector that is related to determining 17 HIV-1 proteins. However, they have gained better performance through incorporating likely interactions (called "partially labeled"), which do not have sufficient evidence to be categorized as direct interaction. The same classifier is used as a quality control in Wuchty (2011), where a RF classifier assesses the quality of candidate interactions, obtained by discovering homologous and

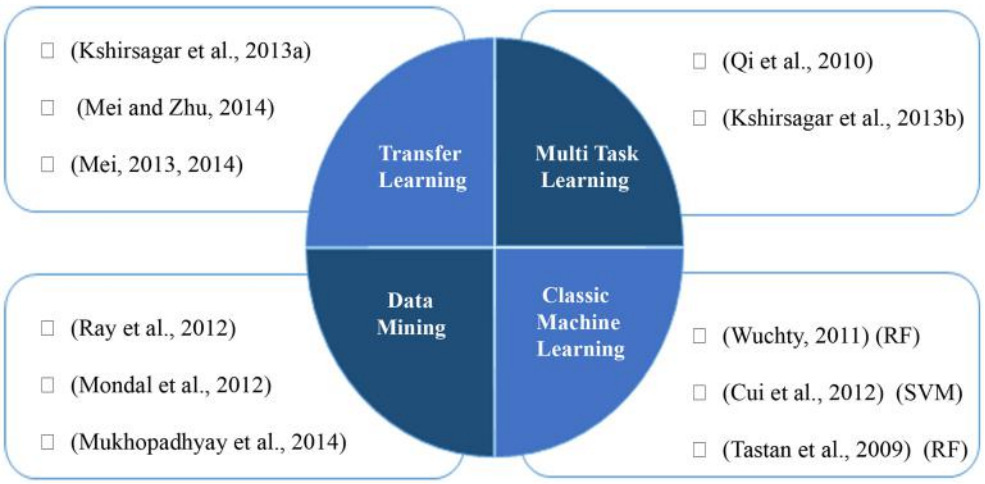


FIGURE 1 | Machine learning and data mining based approaches for prediction of PHIs.

conserved interactions. The author filters the predicted results based on expression and molecular properties.

Conformal prediction is used in Nouretdinov et al. (2012) and the results are compared with those of Tastan et al. (2009) to assess the predictions. This method evaluates the conformance of new pairs with interacting pairs using a method called non-conformity measure (NCM) which shows distinction measure of an example regarding others. Their approach also allows the user to determine confidence level for prediction.

SVM based approaches as a famous classifier are successfully applied in PHI prediction studies (Kshirsagar et al., 2013a; Mei, 2013). Cui et al. (2012) presents a SVM based approach, which uses a fixed length feature vector, indicating relative frequency of consecutive amino acids in the protein sequence. We categorize the machine learning and data mining based approaches in **Figure 1**.

TRANSFER AND MULTITASK LEARNING APPROACHES

One of the promising remedies to tackle the problem of data scarcity is eliciting and transferring data from related domains to desired formulation. Multitask learning uses commonalities among different domains and learn problem simultaneously between them within a shared task formulation, which leads to better performance rather conducting learning task on individual domain. A review paper, Xu and Yang (2011) presents some of the studies utilizing this idea in bioinformatics. For PPI prediction, a method was proposed in Xu et al. (2010) which uses collective matrix factorization originally proposed by Singh and Gordon (2008) to transfer knowledge from a relatively dense PPI network called "source" for predicting new PPIs in a sparse target PPI network. Their goal is to predict intra-species pathogen PPIs as target with the aid of human PPIs as source network through defining a similarity matrix to act as a bridge between them. Another study conducts three different individual classifiers on three GO features (molecular functions, cellular localization, and biological processes) on available protein features and at the same time three classifiers on alternative homolog features to exploit transfer learning. An ensemble classifier produces final result using weighting probability outputs of individual classifiers (Mei, 2013). They applied relatively same idea using a multi instance AdaBoost method to transfer homolog feature as the second instance of proteins (Mei, 2014; Mei and Zhu, 2014). A combination of supervised and semi-supervised approaches is proposed by Qi et al. (2010) through multitask learning. Semi-supervised task on partially positive labels is conducted to improve the supervised classification which trains multi-layer perceptron using labeled data. Another multitask formulation is used in Kshirsagar et al. (2013b) to integrate knowledge from different pathogen-host systems to increase the prediction power of the combined model. Each task is formulated as predicting PHI data between each pathogen and its host. To define similarity between tasks and transfer shared knowledge, they assume that similar pathogens tend to target same biological process in human. In other words, "commonality hypothesis" is introduced that assumes pathway membership of human proteins in positive PHIs should be similar between different tasks. To implement this idea, optimization problem is conducted and dissimilarities are penalized in the objective function. They use transfer learning in Kshirsagar et al. (2013a) for the cases where no known interaction is available by exploiting precisely chosen instances from a source task.

DATA MINING BASED APPROACHES

Machine learning based methods which formulate PPI prediction as a classification task use both interacting and non-interacting protein pairs as positive and negative classes, respectively. Constructing negative class is not straightforward due to the fact that there is no experimentally verified non-interacting pair. This has motivated some studies to overcome this problem by removing the need for negative data through using alternative methods (Mukhopadhyay et al., 2010, 2012, 2014; Mondal et al., 2012; Ray et al., 2012). They integrate bi-clustering with association rule mining, utilizing only positive samples to predict virus-human interactions.

UTILIZED FEATURES

Various studies utilize different sets of biological information through data integration to improve the prediction performance. However, it should be noted that making use of a lot of features without enriching training data may lead to over fitting in the model (Mei, 2014). **Table 2** summarizes the utilized features within different studies on PHI prediction, providing all the cataloged feature information is not always possible for all pathogen systems. Furthermore, various features claimed to have different predictive effects in PHI prediction. Outperforming other features was the motivation for some studies to use GO features in PHI prediction (Mei, 2013, 2014) while features extracted from protein sequences, reported as not promising (Yu et al., 2010).

HANDLING MISSING DATA

Applying machine learning methods and specially supervised learning for situations suffer from data scarcity is challenging. Being limited to well-studied pathogen systems like HIV-1 is the consequence of data dependency. Recently, some solutions are proposed to overcome this limitation by offering substituted values for missing data. For instance, in Kshirsagar et al. (2012) two different methods are proposed including information transfer from other species and model-based imputation. First, they rely on homologous proteins data to provide feature values like GO annotations and gene expression data. This contributes a lot and downgrades the missing data significantly. However, for proteins with no available homolog, they have modeled gene expression value distribution. They have compared the proposed "Cross species imputation" with other imputation techniques. The first method is called "RF" which initiates the missing data to mean value and re-estimate it by choosing the nearest leaf node of the created forest. Another intuitive method is choosing the average of the feature values and the last compared method is discarding any pair with missing value which leads to a reduced dataset. Clear improvements are reported in comparison with the listed imputation methods. It should be noted that using solely statistical methods for estimating features like GO values will be hard due to high dimensionality. Mei (2013) uses homolog information when the features of a protein is unavailable. They have designed various experiments to show the performance of substituting homolog features. Pessimistic experiment, which uses only homolog features to train and test without incorporating any base proteins (called "target" in the article), has promising results, indicating that using homolog information is an effective substitute for the target information to tackle the problem of data unavailability.

THE CHALLENGE OF NON-INTERACTING PPIs

Since there is no available verified non-interacting PPI to be used for training the model, selecting negative data remains as a challenge for PPI prediction. Some studies try to circumvent the obstacle by using methods which do not require negative samples (Ray et al., 2012). However, ignoring non-interacting patterns may increase the rate of false positives (Mei, 2013). The negative set is not defined in Nouretdinov et al. (2012) and instead they use unknown label for other pairs. Most of the studies which formulate the problem as a classification task, have to construct negative class through randomly sampling the data. The rate of positive to negative class is chosen in different manners to avoid biasing classifier toward wrong predictions. A ratio of 1:100 is chosen in Kshirsagar et al. (2012, 2013b) and Tastan et al. (2009) expecting one interaction pair within 100 random pathogen-host pairs. Mei (2013) chooses the same ratio for negative and positive classes, however proposes different idea for choosing negative samples. They put aside sub-cellular co-localized pairs from the negative class and report better performance in comparison with random sampling. The study in Dyer et al. (2011) conducted experiments with different ratios and 10 randomly chosen sets for each ratio and stated that beside clearly different results for different ratios, variability of randomly selected negative samples for each ratio does not have major effect on the result accuracy.

HOMOLOGY BASED APPROACHES

The rationale behind this type of methods is the expectation of conserved interactions between a pair of proteins which have interacting homologs in another species. The conserved interaction is called as "Interolog." The simple method of identifying Interologs is as follows: Consider a template PPI pair (a, b) in a source species, find the homolog a′ in the host and the homolog b′ in the pathogen, conclude that (a′, b′) interact. Simplicity and clear biological basis are the main advantages of these methods. However, homology to known interactions is not sufficient for evaluating the biological evidence of the predicted results. Different filtering techniques should be considered for assessing the feasibility of the interactions under an *in vivo* condition and consequently decreasing the false positives.

A homology detection method using template PPI databases, DIP (Salwinski et al., 2004) and iPfam (Finn et al., 2014), is published in Krishnadev and Srinivasan (2008) to predict PHI pairs. Searching the sequences of host and pathogen proteins within two template databases are conducted to find a superset of all interactions which are physically and structurally compatible. These potential interactions are refined within two additional filtering steps, to detect biologically feasible interactions including integration of expression and sub-cellular localization data. The authors have applied the same procedure for different pathogens in their subsequent works (Tyagi et al., 2009; Krishnadev and Srinivasan, 2011).

Another research uses the conceptually same approach by exploiting sequence similarity augmented with domain-domain interaction detection (Schleker et al., 2012a). They have two compressive reviews of the computational approaches predicting Salmonella-Host interactions (Schleker et al., 2012b, 2014), which include comparing Salmonella-Human and Salmonella-Plant interaction predictions.

Homolog knowledge can be used indirectly as a remedy for data scarcity and data unavailability by homolog knowledge transfer. Mei (2013) uses homolog information (features) when the features of a protein is unavailable. They have designed different experiments to show the performance of substituting homology features. Pessimistic experiment, which uses only homology features for train and test without incorporating any base proteins (called as "target" in the article) has promising results, indicating

Table 2 | Summary of the exploited features for prediction of PHIs.

Utilized feature	Description	References
Domain and motif information	Set to be 1 every domain pair of each PPI in a binary feature vector of all possible domain pairs	Dyer et al., 2011
	Count possible interacting domains between pathogen and host proteins using domain interactions database (3DID)	Kshirsagar et al., 2012, 2013b
	Functional sequence motifs from ELM database checked in HIV-1 sequence	Tastan et al., 2009; Qi et al., 2010; Nouretdinov et al., 2012
	Suppose protein pairs as interacting when they have one or more interacting domain	Coelho et al., 2014
Protein sequence n-mers (n-gram)	For each pathogen-host protein pair concatenate their vectors. Each protein vector count the number of times each distinct n-mer occurred in the sequence	Dyer et al., 2011
	Similar to Dyer et al. (2011)	Kshirsagar et al., 2012, 2013b
	Variant of the spectrum kernel based on sequence n-mers	Kshirsagar et al., 2013a
	Represent proteins by relative count of amino acid 3-mers	Cui et al., 2012
	Forming 7 amino acid classes and computing frequency difference through 343-dimensional vector	Wuchty, 2011
	Forming 4 amino acid classes and computing standardized frequency difference through 64 possible combination	Dong et al., in press
	Observing each of different 20 amino acids within protein sequence	Coelho et al., 2014
Network topology	Two features for each pathogen-host protein pair including human protein's degree and its betweenness centrality	Dyer et al., 2011
	Three features of human protein: degree, clustering coefficient, centrality	Tastan et al., 2009; Qi et al., 2010; Nouretdinov et al., 2012
	Similar to Tastan et al. (2009)	Kshirsagar et al., 2012, 2013b
	Degree and betweenness centrality in human PPI	Dong et al., in press
Gene ontology	Pairwise similarity between GO terms of host and pathogen and Neighbor similarity for GO terms of pathogen and binding partners of human proteins	Kshirsagar et al., 2012, 2013b
	Pairwise and neighbor GO similarity	Tastan et al., 2009; Qi et al., 2010; Nouretdinov et al., 2012
	Three aspects of Gen Ontology are the only used feature values and the homolog GO features are used for missing data	Mei, 2013, 2014
	Biological process similarity is computed for protein pairs	Coelho et al., 2014
	For every human protein within extracted biclusters find important GO terms	Ray et al., 2012; Mukhopadhyay and Maulik, 2014
	Using GO functional data for conducting two functional analysis	Reid and Berriman, 2013
Gene expression	Differential human gene expression infected by pathogen in seven control conditions	Kshirsagar et al., 2012, 2013b

(Continued)

Table 2 | Continued

Utilized feature	Description	References
	Differential human gene expression across HIV-1 infected and uninfected samples	Tastan et al., 2009; Qi et al., 2010; Nouretdinov et al., 2012
Conserved pathways	Find other known PHI, which pathogen is homolog and host proteins share a pathway	Kshirsagar et al., 2012, 2013b
RNAi expression	Utilizing human genes reported as "hits" by the RNAi screens	
Homology information	For each PHI count the number of interologs from other species	
	Forming orthologous groups through clustering host and pathogen proteins around central orthlogous pairs	Wuchty, 2011
	Use STRING to get clusters of orthologous groups and their scores	Coelho et al., 2014
Pfam interactions	Counts the possible interactions between Pfam families of host and pathogen reported in the iPfam	Kshirsagar et al., 2012, 2013b
	Use interacting pair of domains to predict gene interaction between malaria and its hosts (mouse and mosquito)	Reid and Berriman, 2013
Protein sequence	Sequence alignment between pathogen and host proteins computed using PSI-BLAST	Kshirsagar et al., 2012, 2013b
Tissue feature	Check infection susceptibility of tissues	Tastan et al., 2009; Qi et al., 2010; Nouretdinov et al., 2012
Virus protein type	One feature for each HIV-1 protein to compute probability of interacting with human protein	
	A feature vector formed by 11 types of HCV proteins and 9 types of HPV	Cui et al., 2012
Pathways	Pathway participation coefficient is calculated for each protein	Wuchty, 2011
	Use similarity of pathway memberships of human proteins to propose commonality hypothesis across organisms	Kshirsagar et al., 2013b
	For each human protein within extracted biclusters find important KEGG pathways	Ray et al., 2012; Mukhopadhyay and Maulik, 2014
	Find other known PHI, which pathogen is homolog and host proteins share a pathway	Kshirsagar et al., 2012, 2013b

that using homolog information is an effective substitute for the target information to tackle the problem of data unavailability.

Another research uses high confidence intra-species PPIs to detect Interologs using ortholog information (Lee et al., 2008). The assumption is that when two orthologous groups are shared between more than two species, there will be a potential Interolog between those orthologous groups. The potential interactions are filtered using gene ontology annotations followed by pathogen sequence filtering based on the presence or absence of translocational signals to refine the predictions. The notable point is negligible intersection of the predicted interactions with those of the reported predictions in Dyer et al. (2007) due to applying different techniques and datasets for same pathogen-host system.

Zhou et al. (2014) introduces the "stringent homology" which does not rely only on intra-species template PPIs to discover interologs and make use of two different organisms as the source of template PPIs to predict PHIs. They also claim that it is not only for the targeted host proteins which tend to be hub in their own PPI network and this is also true about targeting pathogen proteins.

The most important obstacle for using homology based methods is scarcity of available homolog information. For instance, the number of interologs within bacterial PPIs are not dignificant (Kshirsagar et al., 2013b) demonstrating that we cannot rely only on homolog information for every situation without being cautious about data availability. Clearly, it is reasonable to predict more genomic and proteomic data will be available in the future and consequently more accurate homologs are identified paving the way of studying less-known pathogens. **Table 3** summarizes the published research for predicting PHIs based on homology information.

STRUCTURE BASED APPROACHES

A number of studies are based on structural similarities and use template PPIs to detect similar interacting pairs within host and pathogen proteins. Preliminary ideas presented in Davis et al. (2007) called comparative modeling and was based on their prior work (Davis et al., 2006). Their method starts with a set of host and pathogen proteins and then sequence matching procedures are used to determine the similarities between the

host or pathogen proteins with known structure or known interaction protein partners. Sequence similarity score is only used when structure information is unavailable as a statistical potential assessment, to predict interacting partners. Filtering the set of potential interactions is the last step which is performed using the biological contexts of proteins and a network-level filter. The outcome of this process is decreasing the potential PHIs by about five orders of magnitude. The main drawback of this method is that finding high similarity between pathogen proteins and proteins with known structure is not guaranteed for all pathogen proteins. Therefore, unavailability of the spatial structural information would restrict the applicability of this method. Furthermore, they have only the ability to collect limited number of benchmark PPIs from literature to evaluate their prediction performance.

Authors in Franzosa and Xia (2011) claim to significantly reduce the rate of false positives by presenting virus-human structural interaction network, in which, each PPI is associated with a high confidence 3D structural model. Applicability of the method is limited to human-human and virus-human PPIs for which 3D structural models are available. The method starts with extracting human interacting pairs from PDB and followed by mapping virus proteins to them by sequence similarity. They emphasize the importance of constructing a high-resolution, 3D structural view of pathogen-host and within-host PPI networks to discover new principles of PHIs through their review paper in Franzosa et al. (2012).

Another research developed a map of interactions between HIV-1 and human proteins based on protein structural similarity (Doolittle and Gomez, 2010). A comparison of known crystal structures is performed to measure structural similarity between host and pathogen proteins. Human proteins which have high structural similarity to a HIV protein are identified and their known interacting partners are determined as targets. The assumption is that HIV proteins have the same interactions as their human peers. These predicted results refined by two filtering steps using data from the recent RNAi screens and cellular co-localization information. They apply the same method for developing an interaction network between Dengue virus and its hosts (Doolittle and Gomez, 2011). Again, with a similar idea those proteins with comparable structures share interaction partners. The work suffers from the lack of assessment data in a way that, very limited number of used benchmark PPIs are specific to the viral pathogen. **Table 4** summarizes the conducted research for predicting PHIs based on structural data.

Table 3 | Homology based approaches for prediction of PHIs.

Method	References
Homology detection method using template PPI databases, DIP, and iPfam	Krishnadev and Srinivasan, 2008
Interologs were inferred from ortholog information obtained from high confidence databases	Lee et al., 2008
Homology detection method using template PPI databases, DIP, and iPfam	Tyagi et al., 2009
Homology detection method using template PPI databases, DIP, and iPfam	Krishnadev and Srinivasan, 2011
Introduce stringent homology which uses inter species template PPI	Zhou et al., 2014
Conserved PHI network is generated using interacting proteins of the common conserved inter-species bacterial PPI	Barh et al., 2013
Obtain host-pathogen interactome using sequence and interacting domain similarity to known PPIs	Schleker et al., 2012a
Interolog and Domain based approaches are used to predict PHIs	Li et al., 2012
The ortholog information for the four species are integrated from different databases and interspecies PPI network is constructed followed by dynamic modeling of regulatory responses leads to identifying interactions	Wang et al., 2013

DOMAIN AND MOTIF BASED APPROACHES

The idea of exploiting domains as building blocks of proteins for predicting PPIs is well-studied for single organisms (Wojcik and Schächter, 2001; Pagel et al., 2004) regarding the fact that domains are the mediators of interactions. The approach presented in Dyer et al. (2007) is one of the pioneer published research for predicting PHIs. However, small list of interactions are presented and their biological relevance are not strongly evaluated. To predict interactions between host and pathogen proteins, they present an algorithm that integrates protein domain profiles with interactions between proteins from the same organism. For every pair of functional domains (d, e) which is present in protein pair (g, h) respectively, the probability of interacting (g, h) is assessed using Bayesian statistics. To apply this idea to a pathogen-host system, they identify domains in every host and pathogen proteins and compute the interaction probability for each pair of host and pathogen proteins that contain at least one domain. Assuming M_g as the set of domains contained in protein g the interaction probability of proteins (g, h) is computed as:

$$P(g, h) = 1 - \prod_{d \in M_g} \prod_{e \in M_h} (1 - P(g, h|d, e))$$

The authors have published another study which uses domain profiles as features in supervised machine learning for predicting interactions in HIV-Human system.

Table 4 | Structure based approaches for prediction of PHIs.

Method	References
Comparative modeling of 3D structures	Davis et al., 2007
Sharing interacting partners of structurally similar human proteins to HIV proteins	Doolittle and Gomez, 2010
Structural similarity of Denv proteins to human proteins having known interactions	Doolittle and Gomez, 2011
3D structural interaction network of host-pathogen and within-host PPI networks	Franzosa and Xia, 2011
Assumes that structurally homologous proteins have probably interactors in common	De Chassey et al., 2013

A similar knowledge source is chosen in Kim et al. (2007) which makes use of domain information from InterProScan (Quevillon et al., 2005). They predict PPIs using PreDIN (Kim et al., 2002) and PreSPI (Han et al., 2004) algorithms based on domain information. A study for prediction of interacting proteins of rice and *Xanthomonas oryzae* pathovar *oryzae* (Xoo) also uses domain information (Kim et al., 2008). They presented XooNET which provides about 3500 possible interaction pairs as well as the graphical visualizations of the interaction networks.

The work in Arnold et al. (2012) presents a method to predict and rank bacteria-human PPIs based on domain-domain interactions. They collect a list of Pfam domains and bacterial-human proteins which contains one of the listed domains. Then the data was searched for experimentally verified effectors or their homologs in another bacteria. The result is the possible interactions between Salmonella effectors and host proteins.

Not all pathogen systems are appropriate for applying the mentioned domain based approaches, since domains and the related information are not available for all pathogens. For instance, information on domains and the related statistics are not available for a considerable number of the HIV-1 proteins. Regarding this limitation, the work in Evans et al. (2009) concentrates on protein interactions based on short eukaryotic linear motifs (ELMs) for HIV-1 proteins interacting with human protein counter domains (CDs). They do not accept the idea of having relatively weak link among motif/domain bindings and the actual virus-host PPIs which is presented in Tastan et al. (2009). They predict two kinds of interactions for each virus protein, including direct human protein targets (called H1) which bind to virus via a human CD and a virus ELM and the second type includes indirect interactions in which, host proteins that their normal interactions with H1 proteins are potentially disrupted by competition with an HIV-1 protein. **Table 5** summarizes the conducted research for predicting PHIs based on domain and motif knowledge.

Table 5 | Domain and motif based approaches for prediction of PHIs.

Method	References
PreDIN and PreSPI algorithms based on domain information	Kim et al., 2007
Estimating PPI probability using combining interaction probability of domains	Dyer et al., 2007
XooNET uses Structural Interactome MAP (PSIMAP), Protein	Kim et al., 2008
Experimental Interactome MAP (PEIMAP) and Domain-Domain interactions from iPfam	
Based on ELMs on HIV-1 proteins interacting with human protein counter domains (CDs)	Evans et al., 2009
Predict and rank bacteria-human PPIs based on domain-domain interaction	Arnold et al., 2012
Build the virus-host interactomes by identifying domain interactions between virus and host PPIs followed by topological and functional analysis of the network	Zheng et al., 2014
The viral-human interaction network is modeled based on motif-domain interactions	Segura-Cabrera et al., 2013

PERFORMANCE EVALUATION

The lack of gold standard PHI data and the complexity of PHI mechanisms lead to a hard assessment phase, in a way that predicted interactions are rarely supported by a biological basis. Some studies validate their results by measuring the shared interactions with other published materials (Mukhopadhyay et al., 2012, 2014; Segura-Cabrera et al., 2013). Here we focus on computational metrics which are widely used in publications to evaluate the accuracy of their results, which are shown in **Table 6**.

CONCLUSIONS

Inter-species PPI predictions have gained more popularity in recent years. Computational methods may have important roles in paving the way for experimental PHI verifications by highlighting the high potential interactions and limiting the experimental scope which lead to expense reduction and probably the rapid knowledge development. In this paper, we reviewed the studies which directly focused on computationally PHI prediction. Published approaches are categorized based on pathogen-host and the method they utilize. Clearly some pathogen systems are well studied and targeted in more research regarding the availability of the required data. HIV-1 is the most distinguished pathogen which studied specifically using data-requiring machine learning methods. Therefore, the most important challenge for computationally prediction of PHIs, is the lack of available verified interactions and the relevant feature information in most of the pathogens systems. Data unavailability and scarcity refer to verified interacting PPIs, lack of verified non-interacting protein pairs and missing feature information for proteins. Recent studies have found a new source of data to overcome these limitations. Knowledge transfer from related pathogen systems has shown to be an effective remedy, even for situations with no available interactions. These methods enlighten a promising future direction for establishing computational methods which are augmented with additional transferred knowledge.

Table 6 | Popular evaluation metrics used for PHI prediction.

Metric	Formula	References
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	Cui et al., 2012
Specificity	$\frac{TN}{TN+FP}$	Cui et al., 2012
Sensitivity (Recall)	$\frac{TP}{TP+FN}$	Dyer et al., 2011; Cui et al., 2012
Precision	$\frac{TP}{TP+FP}$	Dyer et al., 2011
F1 score	$\frac{2 * Precision * Recall}{Precision + Recall}$	Kshirsagar et al., 2012, 2013b; Mei, 2013; Coelho et al., 2014
AUC	The area under the ROC curve	Davis et al., 2007; Mei, 2013; Coelho et al., 2014

TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative.

REFERENCES

Aloy, P., and Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics (Oxford, England)* 19, 161–162. doi: 10.1093/bioinformatics/19.1.161

Arnold, R., Boonen, K., Sun, M. G. F., and Kim, P. M. (2012). Computational analysis of interactomes: current and future perspectives for bioinformatics approaches to model the host-pathogen interaction space. *Methods* 57, 508–518. doi: 10.1016/j.ymeth.2012.06.011

Baldi, P., and Brunak, S. (2001). *Bioinformatics: the Machine Learning Approach*. Cambridge: MIT press.

Barh, D., Gupta, K., Jain, N., Khatri, G., León-Sicairos, N., Canizalez-Roman, A., et al. (2013). Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in *C. pseudotuberculosis*, *C. diphtheriae*, *M. tuberculosis*, *C. ulcerans*, *Y. pestis*, and *E. coli* targeted by Piper betel compounds. *Integr. Biol.* 5, 495–509. doi: 10.1039/c2ib20206a

Bock, J. R., and Gough, D. A. (2001). Predicting protein–protein interactions from primary structure. *Bioinformatics* 17, 455–460. doi: 10.1093/bioinformatics/17.5.455

Calderone, A., Licata, L., and Cesareni, G. (2014). VirusMentha: a new resource for virus-host protein interactions. *Nucleic Acids Res.* 43, D588–D592. doi: 10.1093/nar/gku830

Chen, K.-C., Wang, T.-Y., and Chan, C. (2012). Associations between HIV and human pathways revealed by protein-protein interactions and correlated gene expression profiles. *PloS ONE* 7:e34240. doi: 10.1371/journal.pone.0034240

Coelho, E. D., Arrais, J. P., Matos, S., Pereira, C., Rosa, N., Correia, M. J., et al. (2014). Computational prediction of the human-microbial oral interactome. *BMC Syst. Biol.* 8, 24. doi: 10.1186/1752-0509-8-24

Cui, G., Fang, C., and Han, K. (2012). Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics* 13 (Suppl. 7): S5. doi: 10.1186/1471-2105-13-S7-S5

Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H., and Sali, A. (2007). Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 16, 2585–2596. doi: 10.1110/ps.073228407

Davis, F. P., Braberg, H., Shen, M.-Y., Pieper, U., Sali, A., and Madhusudhan, M. S. (2006). Protein complex compositions predicted by structural similarity. *Nucleic Acids Res.* 34, 2943–2952. doi: 10.1093/nar/gkl353

De Chassey, B., Meyniel-Schicklin, L., Aublin-Gex, A., Navratil, V., Chantier, T., André, P., et al. (2013). Structure homology and interaction redundancy for discovering virus-host protein interactions. *EMBO Rep.* 14, 938–944. doi: 10.1038/embor.2013.130

Dong, Y., Kuang, Q., Dai, X., Li, R., Wu, Y., Leng, W., et al. (in press). Improving the understanding of pathogenesis of human papillomavirus 16 via mapping protein-protein interaction network. *Biomed Res. Int.* 890381. doi: 10.1155/2014/890381

Doolittle, J. M., and Gomez, S. M. (2010). Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*. *Virol. J.* 7:82. doi: 10.1186/1743-422X-7-82

Doolittle, J. M., and Gomez, S. M. (2011). Mapping protein interactions between Dengue virus and its human and insect hosts. *PLoS Negl. Trop. Dis.* 5:e954. doi: 10.1371/journal.pntd.0000954

Durmuş Tekir, S., Cakir, T., and Ulgen, K. Ö. (2012). Infection strategies of bacterial and viral pathogens through pathogen-human protein-protein interactions. *Front. Microbiol.* 3:46. doi: 10.3389/fmicb.2012.00046

Durmuş Tekir, S., Çakır, T., Ardiç, E., Sayılırbaş, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen-host interaction search tool. *Bioinformatics (Oxford, England)* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Dyer, M., Murali, T. M., and Sobral, B. W. (2011). Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect. Genet. Evol.* 11, 917–923. doi: 10.1016/j.meegid.2011.02.022

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics (Oxford, England)* 23, i159–i166. doi: 10.1093/bioinformatics/btm208

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2008). The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathogens* 4:e32. doi: 10.1371/journal.ppat.0040032

Dyer, M. D., Neff, C., Dufford, M., Rivera, C. G., Shattuck, D., Bassaganya-Riera, J., et al. (2010). The human-bacterial pathogen protein interaction networks of *Bacillus anthracis*, *Francisella tularensis*, and *Yersinia pestis*. *PloS ONE* 5:e12089. doi: 10.1371/journal.pone.0012089

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics* 2:27. doi: 10.1186/1755-8794-2-27

Finn, R. D., Miller, B. L., Clements, J., and Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res.* 42, D364–D373. doi: 10.1093/nar/gkt1210

Franzosa, E. A., Garamszegi, S., and Xia, Y. (2012). Toward a three-dimensional view of protein networks between species. *Front. Microbiol.* 3:428. doi: 10.3389/fmicb.2012.00428

Franzosa, E. A., and Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U.S.A.* 108, 10538–10543. doi: 10.1073/pnas.1101440108

Han, D.-S., Kim, H.-S., Jang, W.-H., Lee, S.-D., and Suh, J.-K. (2004). PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res.* 32, 6312–6320. doi: 10.1093/nar/gkh972

Kim, J.-G., Park, D., Kim, B.-C., Cho, S.-W., Kim, Y. T., Park, Y.-J., et al. (2008). Predicting the interactome of *Xanthomonas oryzae* pathovar *oryzae* for target selection and DB service. *BMC Bioinformatics* 9:41. doi: 10.1186/1471-2105-9-41

Kim, W. K., Kim, K., Lee, E., Marcotte, E. M., Kim, H., and Suh, J. (2007). Identification of disease specific protein interactions between the gastric cancer causing pathogen, *H. pylori*, and Human Hosts using protein network modeling and gene chip analysis. *Gastric Cancer* 1, 179–187.

Kim, W. K., Park, J., and Suh, J. K. (2002). Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform.* 13, 42–50.

Krishnadev, O., and Srinivasan, N. (2008). A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite. *In Silico Biol.* 8, 235–250. doi: 10.1016/j.ijbiomac.2011.01.030

Krishnadev, O., and Srinivasan, N. (2011). Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol. Macromol.* 48, 613–619. doi: 10.1016/j.ijbiomac.2011.01.030

Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2012). Techniques to cope with missing data in host-pathogen protein interaction prediction. *Bioinformatics* 28, i466–i472. doi: 10.1093/bioinformatics/bts375

Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2013a). "Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks," in *A Workshop at the Annual Conference on Neural Information Processing Systems (NIPS 2013), NIPSWorkshop on Machine Learning for Computational Biology* (Lake Tahoe, NV), 3–6.

Kshirsagar, M., Carbonell, J., and Klein-Seetharaman, J. (2013b). Multitask learning for host-pathogen protein interactions. *Bioinformatics* 29, i217–i226. doi: 10.1093/bioinformatics/btt245

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinformatics* 11 (Suppl. 6):S16. doi: 10.1186/1471-2105-11-S6-S16

Lee, S., Chan, C., Tsai, C., and Lai, J. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* 9:S11. doi: 10.1186/1471-2105-9-S12-S11

Li, Z.-G., He, F., Zhang, Z., and Peng, Y.-L. (2012). Prediction of protein-protein interactions between *Ralstonia solanacearum* and *Arabidopsis thaliana*. *Amino Acids* 42, 2363–2371. doi: 10.1007/s00726-011-0978-z

Mei, S. (2013). Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins. *PloS ONE* 8:e79606. doi: 10.1371/journal.pone.0079606

Mei, S. (2014). Computational reconstruction of proteome-wide protein interaction networks between HTLV retroviruses and *Homo sapiens*. *BMC Bioinformatics* 15:245. doi: 10.1186/1471-2105-15-245

Mei, S., and Zhu, H. (2014). AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. *PloS ONE* 9:e110488. doi: 10.1371/journal.pone.0110488

Mondal, K. C., Pasquier, N., Mukhopadhyay, A., Pereira, C., Maulik, U., and Tettamanzi, A. G. B. (2012). "Prediction of protein interactions on HIV-1–human Ppi data using a novel closure-based integrated approach," in

Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (Vilamoura), 164–173.

Mukhopadhyay, A., and Maulik, U. (2014). Network-based study reveals potential infection pathways of hepatitis-C leading to various diseases. *PloS ONE* 9:e94029. doi: 10.1371/journal.pone.0094029

Mukhopadhyay, A., Maulik, U., and Bandyopadhyay, S. (2012). A novel biclustering approach to association rule mining for predicting HIV-1-human protein interactions. *PloS ONE* 7:e32289. doi: 10.1371/journal.pone.0032289

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., and Eils, R. (2010). "Mining association rules from HIV-human protein interactions," in *2010 International Conference on Systems in Medicine and Biology* (Kharagpur), 344–348. doi: 10.1109/ICSMB.2010.5735401

Mukhopadhyay, A., Ray, S., and Maulik, U. (2014). Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. *BMC Bioinformatics* 15:26. doi: 10.1186/1471-2105-15-26

Navratil, V., de Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., André, P., et al. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res.* 37, D661–D668. doi: 10.1093/nar/gkn794

Nouretdinov, I., Gammerman, A., Qi, Y., Klein-Seetharaman, J., and Learning, C. (2012). Determining confidence of predicted interactions between HIV-1 and human proteins using conformal method. *Pac. Symp. Biocomput.* 311, 311–322. doi: 10.1142/9789814366496_0030

Pagel, P., Wong, P., and Frishman, D. (2004). A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.* 344, 1331–1346. doi: 10.1016/j.jmb.2004.10.019

Qi, Y., Tastan, O., Carbonell, J. G., Klein-Seetharaman, J., and Weston, J. (2010). Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. *Bioinformatics* 26, i645–i652. doi: 10.1093/bioinformatics/btq394

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., et al. (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120. doi: 10.1093/nar/gki442

Ray, S., Mukhopadhyay, A., and Maulik, U. (2012). "Predicting annotated HIV-1 – human PPIs using a biclustering approach to association rule mining," in *2012 Third International Conference on Emerging Applications of Information Technology (EAIT)* (Kolkata), 3–6.

Reid, A. J., and Berriman, M. (2013). Genes involved in host – parasite interactions can be revealed by their correlated expression. *Nucleic Acids Res.* 41, 1508–1518. doi: 10.1093/nar/gks1340

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi: 10.1093/nar/gkh086

Schleker, S., Garcia-Garcia, J., Klein-Seetharaman, J., and Oliva, B. (2012a). Prediction and comparison of Salmonella-human and Salmonella-Arabidopsis interactomes. *Chem. Biodiver.* 9, 991–1018. doi: 10.1002/cbdv.201100392

Schleker, S., Kshirsagar, M., and Klein-seetharaman, J. (2014). Comparing human-Salmonella with plant-Salmonella proteinprotein interaction predictions. *Front. Microbiol.* 5:552. doi: 10.3389/fmicb.2014.00552

Schleker, S., Sun, J., Raghavan, B., Srnec, M., Müller, N., Koepfinger, M., et al. (2012b). The current Salmonella – host interactome. *Proteomics Clin. Appl.* 6, 117–133. doi: 10.1002/prca.201100083

Schleker, S., and Trilling, M. (2013). Data-warehousing of protein-protein interactions indicates that pathogens preferentially target hub and bottleneck proteins. *Front. Microbiol.* 4:51. doi: 10.3389/fmicb.2013.00051

Segura-Cabrera, A., García-Pérez, C. A., Guo, X., and Rodríguez-Pérez, M. A. (2013). A viral-human interactome based on structural motif-domain interactions captures the human infectome. *PloS ONE* 8:e71526. doi: 10.1371/journal.pone.0071526

Singh, A. P., and Gordon, G. J. (2008). "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Lasvegas, NV), 650–658. doi: 10.1145/1401890.1401969

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. *Pac. Symp. Biocomput.* 14, 516–527.

Tyagi, N., Krishnadev, O., and Srinivasan, N. (2009). Prediction of protein-protein interactions between Helicobacter pylori and a human host. *Mol. Biosyst.* 5, 1630–1635. doi: 10.1039/b906543c

Wang, Y.-C., Lin, C., Chuang, M.-T., Hsieh, W.-P., Lan, C.-Y., Chuang, Y.-J., et al. (2013). Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a Candida albicans-zebrafish interaction study. *BMC Syst. Biol.* 7:79. doi: 10.1186/1752-0509-7-79

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Wojcik, J., and Schächter, V. (2001). Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17, S296–S305. doi: 10.1093/bioinformatics/17.suppl_1.S296

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *PloS ONE* 6:e26960. doi: 10.1371/journal.pone.0026960

Xu, Q., Xiang, E. W., and Yang, Q. (2010). "Protein-protein interaction prediction via collective matrix factorization," in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Hong Kong), 62–67. doi: 10.1109/BIBM.2010.5706537

Xu, Q., and Yang, Q. (2011). A survey of transfer and multitask learning in bioinformatics. *J. Comput. Sci. Eng.* 5, 257–268. doi: 10.5626/JCSE.2011.5.3.257

Yu, J., Guo, M., Needham, C. J., Huang, Y., Cai, L., and Westhead, D. R. (2010). Simple sequence-based kernels do not predict protein-protein interactions. *Bioinformatics* 26, 2610–2614. doi: 10.1093/bioinformatics/btq483

Zheng, L.-L., Li, C., Ping, J., Zhou, Y., Li, Y., and Hao, P. (2014). The domain landscape of virus-host interactomes. *Biomed Res Int.* 2014:867235. doi: 10.1155/2014/867235

Zhou, H., Gao, S., Nguyen, N., Fan, M., and Jin, J. (2014). Stringent homology-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *Biol. Dir.* 9, 1–30. doi: 10.1186/1745-6150-9-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 December 2014; accepted: 26 January 2015; published online: 24 February 2015.

Citation: Nourani E, Khunjush F and Durmuş S (2015) Computational approaches for prediction of pathogen-host protein-protein interactions. Front. Microbiol. ***6****:94. doi: 10.3389/fmicb.2015.00094*

This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 04 August 2015
doi: 10.3389/fmicb.2015.00764

Integrated inference and evaluation of host–fungi interaction networks

*Christian W. Remmele[1], Christian H. Luther[1], Johannes Balkenhol[1], Thomas Dandekar[1], Tobias Müller[1] and Marcus T. Dittrich[1,2]**

[1] Department of Bioinformatics, University of Würzburg, Würzburg, Germany, [2] Department of Human Genetics, University of Würzburg, Würzburg, Germany

Fungal microorganisms frequently lead to life-threatening infections. Within this group of pathogens, the commensal *Candida albicans* and the filamentous fungus *Aspergillus fumigatus* are by far the most important causes of invasive mycoses in Europe. A key capability for host invasion and immune response evasion are specific molecular interactions between the fungal pathogen and its human host. Experimentally validated knowledge about these crucial interactions is rare in literature and even specialized host–pathogen databases mainly focus on bacterial and viral interactions whereas information on fungi is still sparse. To establish large-scale host–fungi interaction networks on a systems biology scale, we develop an extended inference approach based on protein orthology and data on gene functions. Using human and yeast intraspecies networks as template, we derive a large network of pathogen–host interactions (PHI). Rigorous filtering and refinement steps based on cellular localization and pathogenicity information of predicted interactors yield a primary scaffold of fungi–human and fungi–mouse interaction networks. Specific enrichment of known pathogenicity-relevant genes indicates the biological relevance of the predicted PHI. A detailed inspection of functionally relevant subnetworks reveals novel host–fungal interaction candidates such as the *Candida* virulence factor PLB1 and the anti-fungal host protein APP. Our results demonstrate the applicability of interolog-based prediction methods for host–fungi interactions and underline the importance of filtering and refinement steps to attain biologically more relevant interactions. This integrated network framework can serve as a basis for future analyses of high-throughput host–fungi transcriptome and proteome data.

Keywords: pathogen–host interaction (PHI), protein–protein interaction, interolog, *Candida albicans*, *Aspergillus fumigatus*, network inference, pathogenicity, bioinformatics and computational biology

OPEN ACCESS

Edited by:
Reinhard Guthke, Leibniz-Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Germany

Reviewed by:
Sven Krappmann, Friedrich-Alexander-Universität Erlangen-Nürnberg – University Hospital Erlangen, Germany
Mangesh Bhide, University of Veterinary Medicine and Pharmacy in Košice, Slovakia

****Correspondence:***
Marcus T. Dittrich, Department of Bioinformatics, University of Würzburg, Am Hubland, Würzburg 97074, Germany
marcus.dittrich@biozentrum.uni-wuerzburg.de

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *04 March 2015*
Accepted: *13 July 2015*
Published: *04 August 2015*

Citation:
Remmele CW, Luther CH, Balkenhol J, Dandekar T, Müller T and Dittrich MT (2015) Integrated inference and evaluation of host–fungi interaction networks. Front. Microbiol. 6:764. doi: 10.3389/fmicb.2015.00764

Introduction

Fungal pathogens infect hundreds of millions of people world-wide every year (Havlickova et al., 2008). Although, the death toll of fungal diseases is comparable to that of malaria or tuberculosis the global burden imposed by fungal pathogens still remains underestimated (Brown et al., 2012). In general, infections caused by fungal pathogens can lead to a diverse range of diseases ranging from superficial infections to invasive mycoses. The outcome of fungal infections is often associated with the intactness of the patients' immune system and therefore fungi pose an increasingly severe threat to the growing numbers of immunocompromised patients in modern medicine, with high mortality rates exceeding 50% for invasive fungal diseases (Brown et al., 2012).

Among fungal pathogens the dimorphic yeast *Candida albicans* and the filamentous fungus *Aspergillus fumigatus* are the most important causes of life-threatening invasive mycoses (Horn et al., 2012). *C. albicans* colonizes the skin and intestinal mucosa of 30–70% of healthy individuals and invasive infection almost exclusively begins endogenously starting from a usually harmless surface colonization, frequently in the gastrointestinal tract (Gow et al., 2012). In contrast to the endogenous pathway of *C. albicans*, infections by *A. fumigatus* mainly occur exogenously via the inhalation of fungal spores (conidia) causing chronic pulmonary aspergillosis or invasive aspergillosis in patients with a severely weakened immune system (Brown et al., 2012). Despite these differences during the infection process, several common strategies of pathogenesis are shared between both fungi.

Host–fungi interactions have been described as commensalism, symbiosis, or pathogenicity. Interestingly, the mechanisms of symbiosis and pathogenicity share common features and there is evidence for parallel trends in evolution between host and pathogens (Ochman and Moran, 2001). The transition from commensal to pathogen is often dependent on small differences (Martin and Nehls, 2009) and the host–pathogen relation can change by environmental conditions (Hube, 2004). Strong adhesion of the fungi to the surface forming a protective biofilm is important for invasive growth, as invasion is driven by pressure on the solid substrate (de Groot et al., 2013). In this sense host–fungal interaction can be characterized by the formation of symbiotic or pathogenic interfaces (Bonfante and Genre, 2010). This relates in particular to processes of pathogen–host interaction (PHI) where both fungi mainly need to overcome similar epithelial barriers and develop skills for the evasion of the innate immune system, capabilities which contribute to the aggressiveness of both pathogens (Horn et al., 2012).

Therefore, a principal aim of systems biological research of human–pathogenic fungi is to unravel the intricate network of interactions between host and the fungal pathogen and elucidate the complex pathogenesis processes of fungal infections. A major quest in this field is the identification of physical or direct interactions between fungus and host proteins during the infection processes. Albeit the research of host–pathogen interactions is becoming increasingly popular in experimental as well as computational science, only a small number of interactions between fungi and human have been reported in literature so far. This leaves a large gap for novel bioinformatical strategies for the prediction of PHI of pathogenic fungi.

With the advent of large scale interaction detection methods the experimental and computational analyses of protein–protein interactions (PPIs) have established an important research field in bioinformatics during the last decade. Still most efforts have been dedicated to the investigation of intraspecies interactions (i.e., interaction between proteins within one species). The primary species in the focus of investigation so far have been *Homo sapiens* and *Saccharomyces cerevisiae*. This is reflected in the fact that the largest experimentally derived PPI datasets available in databases primarily cover *H. sapiens* and *S. cerevisiae* interactions. Currently, these two species constitute almost 74% percent of all non-redundant physical interactions[1] (*H. sapiens*: 50.7% and *S. cerevisiae*: 23.0%) in the BioGRID database (Chatr-Aryamontri et al., 2013). The networks of most other species are considerably smaller and for network analysis these datasets are often extended by the inclusion of interolog based predictions to obtain a larger search space, where interologs are defined as PPIs that are conserved between orthologous proteins in different species (Walhout et al., 2000). Nowadays the interolog approach is commonly used for the classical prediction intraspecies interactions and is particularly valuable for the prediction of novel PPI in species where only a small number of interactions have been experimentally detected. Conceptually, the interactions are transferred from one species to another. This means that if for a given pair of interacting proteins in the source species, homologues for both interaction partners exist in the target species an interaction between those two homologs is inferred. The rational of this interaction transfer is based on the assumption that if a pair of homologous proteins originates from the same ancestral pair of interacting proteins, it can be expected, that the inheritance of the amino acid sequence translates into a related and similar protein structure, and thereby the capability of mutual interaction is also inherited from the ancestral interacting proteins (Walhout et al., 2000). This approach has been extended to the prediction of interspecies interactions and in particular to the prediction of PHIs (Zhou et al., 2013a).

Recent studies investigated the interaction between *H. sapiens* and *Plasmodium falciparum* (Dyer et al., 2007; Lee et al., 2008; Wuchty, 2011), *H. sapiens* and *Helicobacter pylori* (Tyagi et al., 2009), *H. sapiens* and *E. coli* (Krishnadev and Srinivasan, 2011), *H. sapiens* and *Salmonella enterica* (Krishnadev and Srinivasan, 2011) and *H. sapiens* and *Yersinia pestis* (Krishnadev and Srinivasan, 2011) as well as between *H. sapiens* and *Mycobacterium tuberculosis* (Zhou et al., 2014). Apart from the more frequently investigated protozoan *P. falciparum*, most of these studies focus on the interaction with a bacterial pathogens. Fungal infections have only rarely been researched. A recent study examined the interaction between zebra fish and *Candida* (Chen et al., 2013), however, a systemic investigation of direct host–pathogen-PPI between the fungi either *C. albicans* or *A. fumigatus* and the human host has to our knowledge not be conducted so far.

Here we present an extended interolog-based method for the prediction of fungal–host interactions. We focus on the clinically most relevant fungi, the dimorphic yeast *C. albicans* and the filamentous fungus *A. fumigatus*. In addition to the human host, we also investigate interactions between these fungi and *Mus musculus*, since it is the most frequently used animal model in medical sciences. As basic interolog prediction approaches for cross-species analysis often produce large initial predictions sets, we develop and establish an advanced filtering and selection strategy, to reduce the initial set of raw predictions to a smaller refined set of high quality predictions. To this end, we integrate

[1] wiki.thebiogrid.org/doku.php/statistics

information on cellular localization of the predicted host and pathogen interaction partners and focus on proteins associated with cellular functions with relevance for the infection process. The enrichment of established infection and pathogenicity related genes during these subsequent refinement steps emphasizes the biological relevance of the predicted PHIs, from which we highlight and describe some promising candidate interaction in more detail. By this, we demonstrate the benefit of the interolog-based approach in combination with advanced filtering and refinement steps for prediction fungal-host interactions.

Materials and Methods

Template Intraspecies Interaction Networks

For the host–fungi interaction network inference, the intraspecies interaction data of *S. cerevisiae* and *H. sapiens* were downloaded from the following 14 active partners of the International Molecular Exchange (IMEx) consortium (Orchard et al., 2012):

DIP (Salwinski et al., 2004), IntAct (Orchard et al., 2014), MBInfo[2], MINT (Licata et al., 2012), MatrixDB (Chautard et al., 2011), Molecular Connections[3], I2D (Brown and Jurisica, 2007), InnateDB (Breuer et al., 2013), UCL-BHF group, UCL London[4], UniProt Swiss-Prot group, SIB (The UniProt Consortium, 2014), BioGRID (Chatr-Aryamontri et al., 2013), MPact (Pagel et al., 2005), BIND (Bader et al., 2001), and MPIDB (Goll et al., 2008).

PSICQUIC queries (Aranda et al., 2011) were used to retrieve human and yeast intraspecies interaction information from this databases on 09/09/2014. Non-coding genes, interaction loops of self-interacting proteins as well as interactions of the interaction types "colocalization," "additive genetic interaction defined by inequality," "suppressive genetic interaction defined by inequality," "synthetic genetic interaction defined by inequality," "genetic interaction," "genetic inequality," "genetic interference," and "self-interaction" were not used for the template networks.

Orthology Information

Orthology information for *C. albicans*, *S. cerevisiae*, *H. sapiens*, *M. musculus*, and *A. fumigatus* was downloaded from InParanoid8 (Sonnhammer and Ostlund, 2014). Additionally, orthology relations between *A. fumigatus* and *S. cerevisiae* were retrieved from Aspergillus Genome Database (AspGD; Cerqueira et al., 2014). Orthologies of the species pair *A. fumigatus* and *H. sapiens* which was neither available from InParanoid8 nor AspGD, were computed via the InParanoid version 4.1[5] using parameters comparable to the parameters of similar species pairs (*H. sapiens* – *A. kawachii*). Blast version 2.2.26 with the scoring matrix Blosum62, a score-cutoff of 40 bits, a sequence overlap of 0.5, a group merging cutoff 0.5 and a minimal score of 0.05 was used as InParanoid settings. The dataset for *A. fumigatus* protein sequence was downloaded from AspGD, while the protein sequences of *H. sapiens* originated from the InParanoid8 server.

Gene Ontology

Gene Ontology (GO) slim annotations, a subset of the GO dataset (Ashburner et al., 2000) were used to categorize genes in host–fungi interactions of the inferred networks regarding three domains: biological process, molecular function and cellular component. GO slim associations were retrieved from the Candida Genome Database (CGD; Arnaud et al., 2005) and the AspGD (Cerqueira et al., 2014) for both fungal pathogen species. GO slim associations for the host species (*H. sapiens* and *M. musculus*) were downloaded from EnsEmbl 76 (Flicek et al., 2014).

Genes of the inferred fungi–host interaction networks were categorized by GO slim cellular component annotation in likely and unlikely host–fungal interactors under the assumption that interacting host and fungal proteins have to be localized on potential interface (e.g., cell surface or endosome membrane). The GO slim cellular component terms for likely interspecies interactions on the fungal and host side were listed in **Table 1**.

Similar to the refinement step for protein localization, proteins with pathogenicity-associated GO slim biological process terms were selected to enrich for pathogenicity-relevant interaction predictions (see **Table 2**). Only genes assigned to one of the referenced cellular component and biological process GO terms were used for further analyses.

Gene Ontology and Uniprot Tissue Enrichment

Interactors of subnetworks were tested for enriched GO annotation level 2 terms of the domains "biological process," "cellular component," "molecular function" (Ashburner et al., 2000) versus the GO terms background frequencies of the interactors in the full network. The functional enrichment tests were performed via the DAVID Bioinformatics Resources 6.7 (Huang da et al., 2009a,b) using GO terms of all levels and only reporting groups of the size of least two genes and an EASE Score Threshold (for gene-enrichment analysis modified Fisher Exact P-Value) of less than 0.1. The *p*-values were adjusted for multiple testing (Hochberg and Benjamini, 1990). Similar to the GO enrichment, the tissue enrichment analyses were performed on Uniprot tissue terms via the DAVID Bioinformatics Resources 6.7.

Catalog of Pathogenicity-Relevant Genes

To get a set of genes of *H. sapiens* and *M. musculus* that are known to be involved in host–pathogen interactions, the PPI information were downloaded from the HPIDB version 5/22/2014 and the PATRIC database version Mar2013. Further, all interspecies interactions that involved viral pathogens or the interaction types which are not related to a direct PPI such as annotated as "colocalization," "additive genetic interaction defined by inequality," "suppressive genetic interaction defined by

[2]http://www.mechanobio.info/

[3]http://www.molecularconnections.com

[4]http://www.ucl.ac.uk/functional-gene-annotation/cardiovascular

[5]http://software.sbc.su.se/cgi-bin/request.cgi?project=inparanoid

TABLE 1 | Numbers of genes in the primary predicted host–fungal PPI networks belonging to the cellular component GO filter terms.

(A) Filter terms for host side

GO slim cellular component terms	Number of genes in *Homo sapiens*	Number of genes in *Mus musculus*
Extracellular region	2,566	783
Plasma membrane	2,310	2,024
Extracellular space	631	419
Endosome	476	421
Lysosome	306	247
Cilium	138	202
Proteinaceous extracellular matrix	115	132
External encapsulating structure	3	5
Only other GO terms	6,531	7,645
No GO terms	902	361

(B) Filter terms for fungi side

GO slim cellular component terms	Number of genes in *Aspergillus fumigatus*	Number of genes in *Candida albicans*
Plasma membrane	270	236
Extracellular region	94	33
Cell wall	52	74
Only other GO terms	3214	3160
No GO terms	0	1114

(C) Sizes of host–fungi PPI networks after localization refinement

Host species	Pathogen species	Number of host–pathogen interactions	Number of host interactors	Number of pathogen interactors
H. sapiens	*A. fumigatus*	17,853 (8.4%)	363 (10.2%)	2,393 (21.2%)
H. sapiens	*C. albicans*	15,330 (4.3%)	301 (6.6%)	2,123 (19.2%)
M. musculus	*A. fumigatus*	9,284 (4.5%)	337 (9.4%)	1,572 (14.9%)
M. musculus	*C. albicans*	8,055 (2.4%)	282 (6.2%)	1,376 (13.3%)

inequality," "synthetic genetic interaction defined by inequality," "genetic interaction," or "genetic inequality" were removed from the dataset.

Also, the Victors database of PHIDIAS (Xiang et al., 2007), a database containing virulence factors originating from literature curation and bioinformatics analyses and the PHI-base (Winnenburg et al., 2008), a database containing expertly curated molecular and biological information on pathogenic genes experimentally verified to have an effect on the virulence outcome were searched for genes of the fungal pathogens *A. fumigatus* and *C. albicans* that are known as pathogenesis associated.

Additionally the public available interaction databases mentha (Calderone et al., 2013), HPIDB (Kumar and Nanduri, 2010), APID (Prieto and De Las Rivas, 2006), PHISTO (Durmus Tekir et al., 2013), PRIMOS (Rid et al., 2013), and the databases of IMEx (Orchard et al., 2012) were scanned to receive all already known interspecies interactions for human–*Candida*, human–*Aspergillus*, mouse–*Candida*, and mouse–*Aspergillus*.

To find already known human–*Aspergillus*, mouse–*Aspergillus*, human–*Candida*, and mouse–*Candida* interactions the public available interaction databases mentha, HPIDB, APID, PHISTO, PRIMOS, and the databases of IMEx was searched.

Analysis of Dual RNA-Seq Data

For the comparison of predicted fungal–host interaction networks, gene expression data of a previously published time course of murine bone marrow derived dendritic cells phagocytosing *C. albicans* SC5314 cells was used (Tierney et al., 2012). The gene expression data constitutes of dual RNA-seq data simultaneously measuring the transcripts of *Candida* and mouse cells at 30, 60, 90, and 120 min post-infection. The sequenced reads were downloaded from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-595/. Contamination of poly-T at the read start and poly-A at the read end was removed via cutadapt version 1.6 (Martin, 2011). The curated reads were mapped on a combined reference of the *C. albicans* SC5314 version A22 (Arnaud et al., 2005) and the *M. musculus* version GRCm38.75 (Flicek et al., 2014) genome, using the short read mapping tool STAR version 2.4 (Dobin et al., 2013). For each gene of the *C. albicans* and the *M. musculus*, the uniquely mapped reads were counted with featureCounts version 1.4.3 (Liao et al., 2014). Fungal and host genes were tested for differential expression in the infection time course with DESeq2 version 1.6.2 (Love et al., 2014). Genes were identified as differentially expressed when they showed a significant (p-value <0.05) change in read counts after multiple testing correction (Hochberg and Benjamini, 1990).

TABLE 2 | Numbers of genes in the primary predicted host–fungal PPI networks belonging to the biological process GO filter terms.

(A) Filter terms for host side

GO slim biological process terms	Number of genes in *H. sapiens*	Number of genes in *M. musculus*
Signal transduction	951	602
Immune system process	491	255
Symbiosis, encompassing mutualism through parasitism	260	0
Cell adhesion	151	127
Circulatory system process	50	53
Only Other Slim BP annotations	1,246	878
No GOSlim BP annotation	0	0

(B) Filter terms for fungi side

GO slim biological process terms	Number of genes in *A. fumigatus*	Number of genes in *C. albicans*
Pathogenesis	30	33
Cell adhesion	10	24
Biofilm formation	0	32
Interspecies interaction between organisms	0	30
Growth of unicellular organism as a thread of attached cells	0	2
Only Other Slim BP annotations	330	244
No GOSlim BP annotation	0	0

(C) Sizes of host-fungi networks after functional refinement

Host species	Pathogen species	Number of host–pathogen interactions	Number of host interactors	Number of pathogen interactors
H. sapiens	*A. fumigatus*	1,137 (6.4%)	607 (25.4%)	33 (9.1%)
H. sapiens	*C. albicans*	3,025 (19.7%)	840 (39.6%)	57 (18.9%)
M. musculus	*A. fumigatus*	590 (6.4%)	355 (22.6%)	26 (7.7%)
M. musculus	*C. albicans*	1,462 (18.2%)	461 (33.5%)	41 (14.5%)

Network Visualization

The networks were visualized by Cytoscape (Shannon et al., 2003). The top 10% of fungal high degree interactors were removed from the visualized networks to improve the readability. The GO slim interaction network was based on grouping genes in GO slim groups that are annotated by the respective GO slim biological process terms. Improved readability of GO slim networks was achieved by merging GO slim groups fully contained in larger groups. Node size represents the number of genes contained in each GO slim term. Edge width and color depict number of interactions between two GO slim terms.

Results

Host–Fungi Interaction Data in Literature and Public Databases is Sparse

The primary objective of our work is to establish a comprehensive catalog of host–fungal interactions. A first literature search revealed that overall not much detailed data concerning PHIs for fungi is available so far. However, as PHIs have become an important topic in the last years, several databases for PHIs have been established. Up to date most of the interactions deposited in these databases still relate to viral and bacterial pathogens and almost no information concerning fungi is available at all. For example, the current HPIDB (Kumar and Nanduri, 2010) covers predominantly viral (74%: 29,942) and bacterial (22%: 8,992) pathogens and only 4% (1,628) of the interactions involve fungal species out of which over 92% (1,499) relate to *Saccharomyces* spp. To obtain a comprehensive overview of all host–fungi interaction data available so far, we first searched the content of the most prominent host–pathogen interaction databases [HPIDB, PHISTO (Durmus Tekir et al., 2013), and PRIMOS (Rid et al., 2013)] for established host–fungal interactions between human–*Candida*, human–*Aspergillus*, mouse–*Candida*, and mouse–*Aspergillus*. Nevertheless, the search returned only two distinct interactions between *C. albicans* and *H. sapiens* and one more for mouse–*Candida*: (i) *Candida* ORC1 (Origin recognition complex subunit 1) and human CDC23 (Cell division cycle protein 23), (ii) *Candida* Q00308 and human CD2BP2 (CD2 antigen cytoplasmic tail-binding protein 2). For the interaction between mouse and *Candida* only one interaction between the *Candida* CDC28 (Cyclin-dependent kinase 1) and murine Cdkn1b (Cyclin-dependent kinase inhibitor 1B) could be found. We could not find any interspecies interaction between human and *A. fumigatus* or between mouse and *A. fumigatus* from the above host–pathogen-databases. Therefore, we subsequently scanned APID (Prieto and De Las Rivas, 2006), mentha (Calderone et al., 2013) and all the 14 curated PPI databases of the IMEx consortium (Orchard et al., 2012) for cross-species

interactions involving *A. fumigatus* and *C. albicans* (see Catalog of Pathogenicity-Relevant Genes section). This extended search revealed only one additional interspecies interaction that was not included in the PHI databases: *Candida* CDC42 (Cell division control protein 42 homolog) and the murine Scd2 (Acyl-CoA desaturase 2). No interactions for *A. fumigatus* have been found in above databases for human or mouse.

Since information in databases about PPIs between the fungal pathogens *C. albicans* and *A. fumigatus* and their hosts is sparse, we propose a framework to infer PHIs and thus create hypotheses for experimental validation.

Dual Template Interolog-Based Host–Fungi PPI Network Inference Approach

The general approach applied in this study aims on the identification of novel potential PPIs between the selected host species *H sapiens* and *M. musculus* and the fungal pathogen species *C. albicans* and *A. fumigatus*. To derive these PHIs, we established an interolog-based inference method exploiting known intraspecies interactions in *H. sapiens* and *S. cerevisiae* as template networks combined with gene homology information between the template species and the host as well as the fungal species. Our approach comprises three steps which involve (i) the establishment of a comprehensive dual-species PPI template network, (ii) homology based inference of PHIs, and (iii) the application of an extended filtering strategy on the raw predictions to attain a core set of refined interaction predictions (see **Figure 1**).

Comprehensive Dual-Species PPI Template Network

To establish a comprehensive intraspecies template network for interspecies PHI interaction prediction we screened the BioGRID database (Chatr-Aryamontri et al., 2013) and 13 PPI databases associated with the Imex consortium for intraspecies PPIs in *H. sapiens* and *S. cerevisiae* resulting in 170,774 human interactions with 15,509 interactors and 272,167 yeast interactions with 5,824 interactors. As we primarily focus in this study on direct PPIs, the template networks were curated from PPIs detected by methods which are rather based on functional associations (e.g., "genetic interference"). Furthermore, all self-interactions were removed from this network. The resulting human and yeast intraspecies PPI networks consisted of 147,760 human interactions with 15,240 interactors and 130,665 yeast interactions with 5,789 interactors. Although the numbers of human interactions were reduced by almost 14%, the number of interactors barely decreased (1.7%). Since a large number of yeast

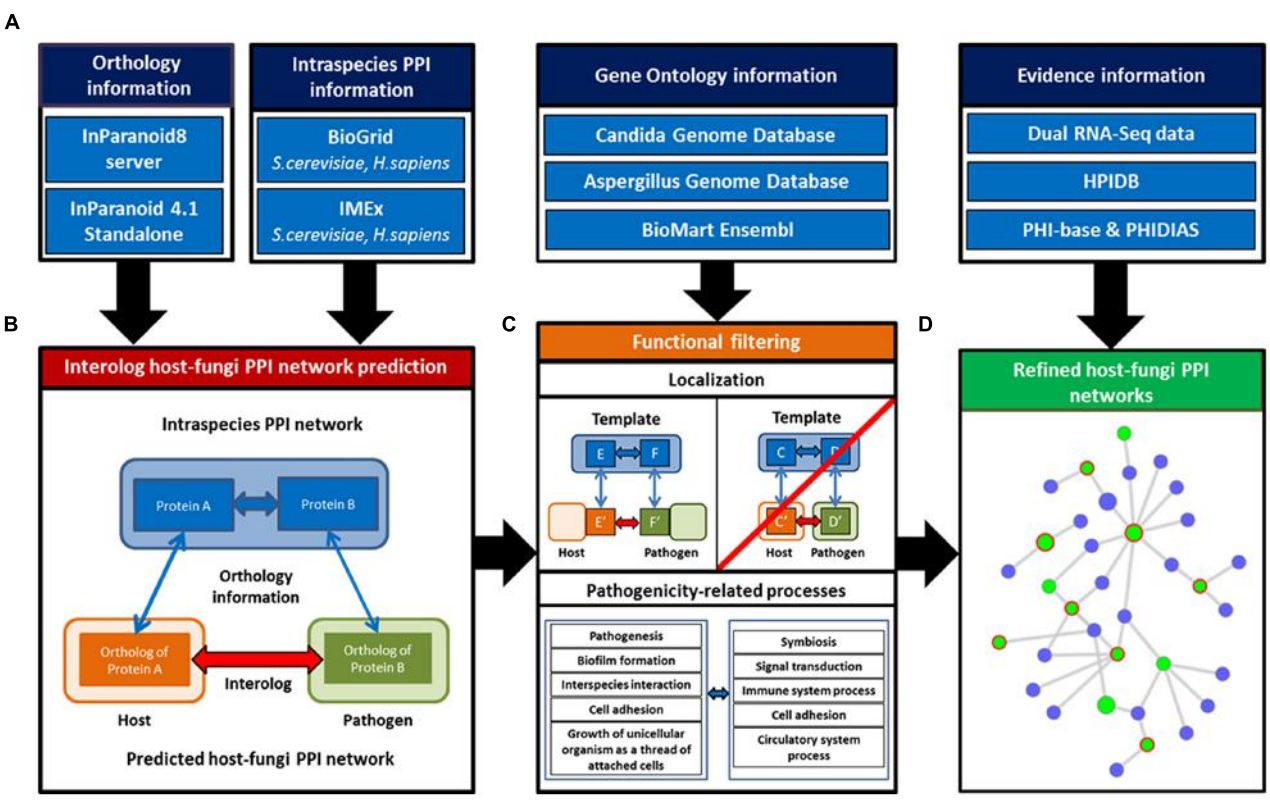


FIGURE 1 | Basic concept of the host–fungi PPI inference and refinement steps. (A) Information of direct PPI from multiple public databases were integrated for the two template networks *Homo sapiens* and *Saccharomyces cerevisiae*. **(B)** These combined with orthology information allowed to identify host–fungi interologs. **(C)** Primary inferred networks were filtered for interactions which showed protein localizations pointing to possible interfaces between host and fungi. Additionally, the networks were refined for pathogenicity-related processes. **(D)** Evidence information of several independent sources (e.g., transcriptome data) were exploited to evaluate the refined host–fungi PPI networks.

interaction were identified by functional association methods, the number of interaction decreased by almost 52%, while similar to the human network the number of interactors was just reduced by less than 1% (see Supplementary Table S2).

Interolog-Based Prediction Yields Large Host Fungal Interaction Networks

Host–fungal interactions for each host–fungi pair were predicted based on the two template interaction networks. Thus, in a second step, we integrated the template interaction data with orthology information of the host, pathogen, and template species. Orthology information between the two template PPI networks of *H. sapiens* and *S. cerevisiae* and the host species *H. sapiens* and *M. musculus* as well as the fungal pathogens *C. albicans* and *A. fumigatus* was downloaded from the InParanoid 8 database (Sonnhammer and Ostlund, 2014), the species-specific genome databases (Binkley et al., 2014; Cerqueira et al., 2014; Costanzo et al., 2014) and missing species pairs complemented by orthology identification by the stand-alone program InParanoid 4.1 (Ostlund et al., 2010). For *H. sapiens* as template species, 16,582 mouse genes were identified as orthologs to 16,417 human genes, while 2,687 *Candida* genes were orthologs to 3,770 *H. sapiens* genes (2,808 *Aspergillus* and 4,277 *H. sapiens* genes, respectively). Interestingly, we found more than twice the number of *Candida* proteins being orthologs to yeast than orthologous *A. fumigatus* proteins, while the number between both fungi and human was comparable to *S. cerevisiae* – *A. fumigatus* orthologs (see Supplementary Table S1)

We searched for orthologs for both interactors of each template interaction to predict potential direct PPIs between the host species *H. sapiens* or *M. musculus* with the fungal pathogen species *C. albicans* or *A. fumigatus*. Interologs are PPIs inferred from one species to another by using orthology information (Walhout et al., 2000). In our approach, we simultaneously identified orthologs of one interactor in the host species and one interactor in the fungal species for each template interaction. The resulting cross-species interologs between the hosts and the pathogens should consequently have the potential to perform a PPI, given both interactors share the same location at one point in time. For the human-*Aspergillus* infection 213,518 interologs with 11,279 human and 3,576 *Aspergillus* interactors could be superimposed. Similar results were obtained for the three other infection setups human–*Candida*, mouse–*Aspergillus*, and mouse–*Candida* (see Supplementary Table S2).

Improving Primary Inferred Host–Fungi PPI Networks

Potential false predictions were reduced via refinement of the primary inferred host–fungi PPI networks based on functional data. Therefore, GO slim annotations of the cellular component and biological process (The Gene Ontology, 2014) were exploited in this filtering step. To enrich for likely interactions, only host and pathogen interactors which showed GO slim cellular component annotations pointing at locations associated to the cell surface and intracellular compartments which can be in direct host–fungi contact, were selected for the refined host–fungi PPI networks. The GO slim cellular compartment terms which were selected for filtering interactors based on their localization were summarized for the hosts (see **Table 1A**) and the fungi (see **Table 1B**). Only 902 human and 361 mouse genes showed no GO slim cellular component annotation at all. On the fungal side, this was the case for 1114 *Candida*, but none of *Aspergillus* genes. Altogether, only very few genes were lost in this filtering step due to missing localization information. The distribution of filtered GO slim cellular component terms clearly shows that the "extracellular region" is less abundant in the murine compared to the human interactor set (783 and 2566), while the other terms are similarly present between mouse and human. Surprisingly, the term "extracellular region" also shows a strong difference in distribution on the fungal side (94 *Aspergillus* interactors and 33 *Candida* interactors).

This filtering step reduced the interolog networks, e.g., human–*Aspergillus* with 213,518 interologs to 17,853 interactions with 2,393 human and 363 *Aspergillus* interactors. For all four interolog networks, the refinement step reduced the number of interactions to less than 9%, while the host interactors were reduced to less than 11% and the fungal interactors to less than 22%, respectively (see **Table 1C**).

In concordance with the localization filtering, a functional refinement utilizing representative biological process terms was applied. To improve the quality of the predicted network and increase the fraction of PPIs potentially associated to pathogenicity-relevant processes, we selected five GO slim biological process terms for filtering the host interactors (see **Table 2A**) and five GO slim biological process terms on the pathogen side (see **Table 2B**). All genes of the hosts and fungal pathogens showed an annotation of GO slim biological process.

In the localization-refined PPI networks, GO slim biological process annotations were available for each host and fungi interactor. Nonetheless, the number of human interactors assigned to the selected GO slim biological process terms was higher than for mouse. Especially, the GO slim term "Symbiosis, encompassing mutualism through parasitism" yielded the strongest difference with a coverage of 260 human interactors and 0 mouse interactors. For the fungal pathogens, the results were similar with fewer *A. fumigatus* interactors than *C. albicans* interactors assigned to selected GO biological process terms.

This filtering step reduced the localization-refined networks, e.g., mouse–*Candida* 8055 interactions with 1,376 mouse and 282 *Candida* interactors to 1,462 interactions with 461 mouse and 41 *Candida* interactors. For all four host–fungi networks, the refinement step reduced the number of interactions to less than 20%, while the host interactors were reduced to less than 40% and the fungal interactors to less than 19%, respectively (see **Table 2C**).

The Dual Template Approach Substantially Enhances the Prediction Space for Host Fungal Network Inference

To investigate the benefits of our dual-template approach for the interolog-based network inference, we examined for each host and fungal interactors the template network from which

they were inferred. For this, we grouped the interactors of the primary inferred PHI networks based on their original template network (see **Figure 2**). On the host side, the human template exclusively makes up for 67.5% of the human interactors in the PHI networks, while over 10.2% of the human interactors originated only from the yeast template (see Supplementary Figure S1). About 22.3% of the human interactors were inferred from both the human and the yeast template. Similarly, for the mouse interactors, the human template solely makes up for over 66.0% of the murine interactors in the PHI networks, while more than 11.5% of the interactors originated only from the yeast template. About 22.4% of the murine interactors were inferred from both the human and the yeast template. Even though no orthology information was required for the inference of human interactors, we see similar distribution of template origin between human and murine interactors. On the fungal side, a substantially larger fraction of the *Aspergillus* interactors (24.4%) was inferred from yeast template, while the human template makes up for 42.4% of the *Aspergillus* interactors originating from the human template. Over 33.1% of the *Aspergillus* interactors were inferred from both the human and the yeast template. In contrast, only less than 8.5% of the *Candida* interactors were inferred from the human template, while more than 43.0% originated from yeast interologs. The largest fraction with more than 48.4% of the *Candida* interactors resulted from both human and yeast template. These numbers represent substantial differences in the distribution between both fungal pathogens, as could be expected by the smaller evolutionary distance from *S. cerevisiae to C. albicans* than from *S. cerevisiae* to *A. fumigatus*.

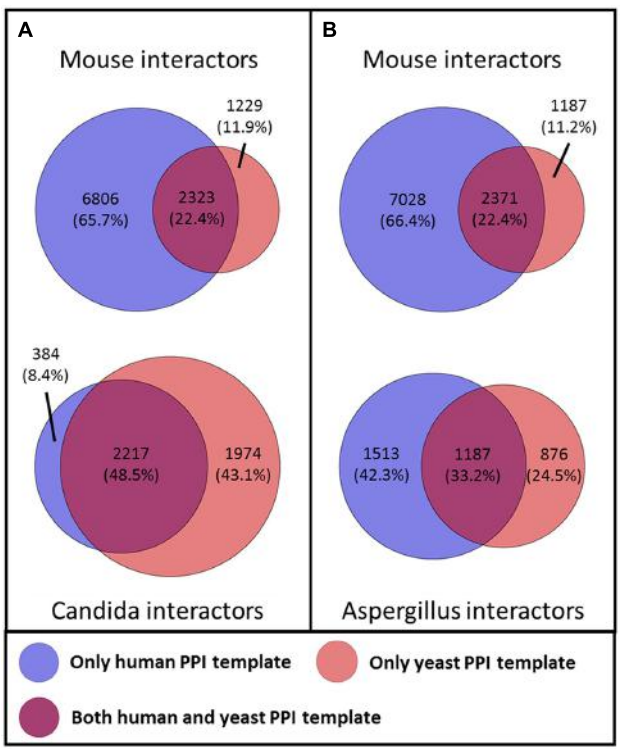


FIGURE 2 | Influence of the template networks on the predicted (A) mouse–*Candida* network (B) mouse–*Aspergillus* network. The color of the circle denotes the template network from which the interactors originated.

A GO enrichment analysis was performed for each group of interactors originating from human, yeast, or both template interaction networks compared to the whole set of interactors (see Supplementary Tables S3 and S4). The GO enrichment analyses showed that multiple GO categories related to PHI were significantly enriched in the human interactor subsets originating from the human template network (e.g., extracellular region part, cell adhesion, signal transducer activity) and yeast template network (e.g., membrane part, transmembrane transport, ion binding). Surprisingly, the subset of human interactors inferred by both template networks was enriched for GO categories of basic biological processes (e.g., intracellular part, ribonucleoprotein complex, nucleotide binding). Even with the overlap of subsets showing only few interesting enriched GO categories, the integration of both template networks complemented a large amount of significantly enriched pathogenicity-relevant categories (see Supplementary Table S3).

Similar to the host side, the GO enrichment analysis of the *Aspergillus* interactors predicted based on the human template network yielded significantly enriched pathogenicity-associated GO terms (e.g., oxidation reduction, ion binding). For the interactors originating from the yeast template network, a different set of pathogen-relevant GO terms (e.g., membrane, transferase activity) were enriched, while the *Aspergillus* interactors inferred by both template networks mainly basic biological processes were enriched (e.g., ribonucleoprotein complex, cellular metabolic process, structural constituent of ribosome; see Supplementary Table S4).

Localization Filtering and Functional Refinement Improve Predicted Host–Fungi Networks

Since data on experimentally validated PHIs for fungal pathogens are rare and there is no golden standard for PHI network inference available, we created a dataset of pathogenicity-associated genes for validation of the refinement step. We extracted functional data encompassing (1) human and murine genes which have been reported to directly interact with pathogenic proteins (Kumar and Nanduri, 2010), (2) virulence and pathogenicity phenotypes induced by knock outs of fungal genes (Xiang et al., 2007; Winnenburg et al., 2008) and (3) infection responsive genes identified by analysis of a data set of an infection time course experiment of murine innate immune cells infected by *C. albicans* (Tierney et al., 2012).

Infection-Regulated Genes are Enriched in Resulting Host–Fungi Networks

Under the assumption, that deregulated genes over an infection time course are more likely to be involved in host–fungi interactions, exploiting transcriptomic or proteomic gene

expression data can be used for the validation of the refinement step. The recently published simultaneous transcriptome sequencing of *C. albicans* and murine innate immune cells 0, 30, 60, 90, and 120 min post-infection uncover the temporal dynamics of infection-regulated genes (Tierney et al., 2012). For 21,251 mouse genes and 6,274 *Candida* genes, we found at least one RNA-seq read matched and performed statistical analyses of all time points compared to 0 min post-infection. This revealed 413 significantly deregulated genes in the mouse transcriptome and 1,068 significantly deregulated genes in the fungal transcriptome. The number of deregulated mouse genes was increasing from time point to time point: 45 genes after 30 min, 169 genes after 60 min, 239 genes after 90 min, and 300 genes after 120 min). Similar to mouse, the number of significant *Candida* genes was also increasing with 314 genes after 30 min, 316 genes after 60 min, 432 genes after 90 min, and 744 genes after 120 min post-infection (see **Figures 3A,B**). Interestingly, significantly deregulated genes in mouse were mainly upregulated genes, at a ratio 5:1. In contrast, the significant genes in *Candida* showed almost the same number of up- and downregulated genes.

With the identified deregulated genes in *C. albicans* and *M. musculus*, we generated a set of infection-associated genes each for the fungal pathogen and the mammalian host. With these sets as a positive list, deregulated genes were significantly enriched in the final refined network compared to the primary inferred mouse–*Candida* PPI network (see **Figures 3C,D**). For the predicted mouse interactors, the localization-based filtering step did not show a significant enrichment in contrast to the functional refinement. Due to the small number of interactors (12 of 41) in the refined network, the functional refinement step did not show a significant enrichment for the predicted *Candida* interactors. While the deregulated mouse genes were significantly enriched by the interolog-based inference step, the significant *Candida* genes were significantly depleted. This showed that for a vast number of pathogen-related genes in *Candida*, there were no interologous interactions found in the template networks.

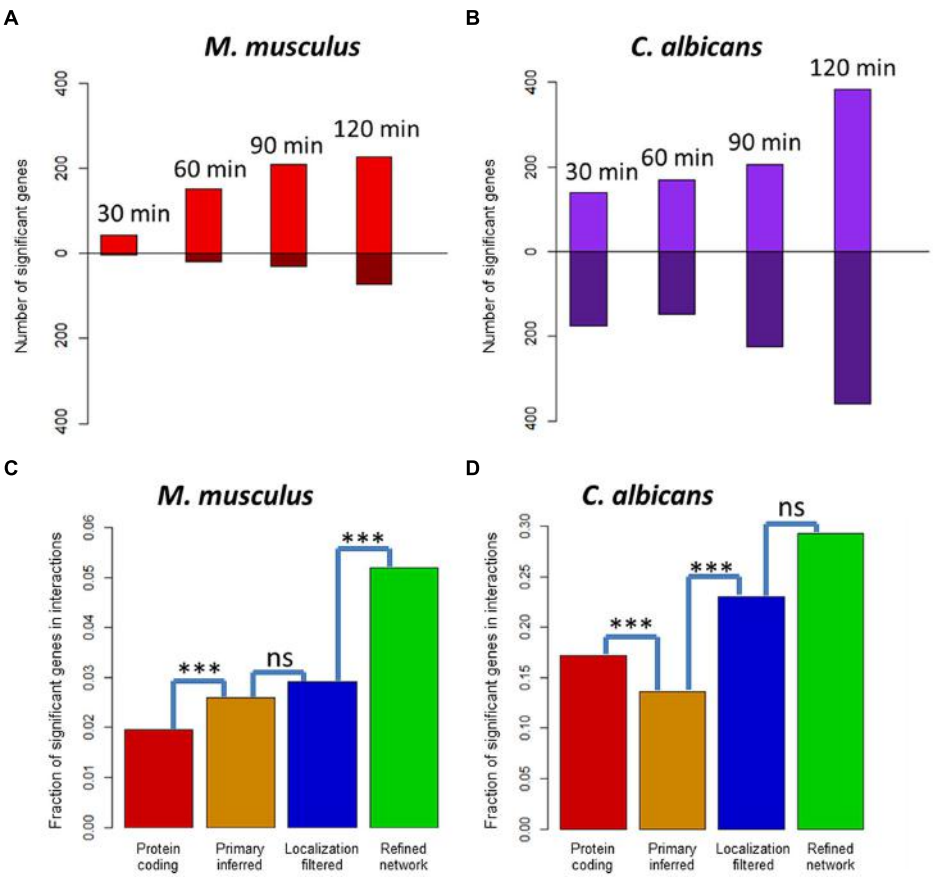


FIGURE 3 | Infection-regulated genes in predicted host and fungi interactors. (A) Differentially expressed genes in murine innate immune cells 30, 60, 90, and 120 min post-infection with *Candida albicans* cells. Bars above the *x*-axis show the number of significantly upregulated genes, while bars below show the significantly downregulated genes. **(B)** Differentially expressed genes in *C. albicans* 30, 60, 90, and 120 min post-infection of murine innate immune cells. **(C)** Fraction of significantly deregulated genes in the sets of protein-coding genes, the primary inferred, the localization-filtered, and the functionally refined interactors of mouse. **(D)** Fraction of significantly deregulated genes in the sets of protein-coding genes, the primary inferred, the localization-filtered, and the biological process refined interactors of *C. albicans*. A test for enrichment of infection-regulated genes in the interactor sets after the primary inference, localization, and functional refinement step (Fisher exact test, $***p < 0.001$).

Pathogenicity-Associated Genes are Enriched in Resulting Host–Fungi Networks

Since databases even specialized on PHI contained very few PPI between human and fungal (mainly *S. cerevisiae*) pathogens [e.g., HPIDB comprised 126 host–fungal PPIs], we extracted all human genes interacting with Archaean (0.03%), protozoan (0.3%), fungal (3.6%), or bacterial (96.1%) pathogen genes. Viral interactions were not included in our dataset as these interactions are mainly intracellular. This yielded pathogenicity-associations for 3,419 of the 20,688 protein-coding human genes which translates to a fraction of 16.5%. In contrast to the large number of human interactors, there were only 32 PHI mouse genes in the database. Because of the small number of mouse genes interacting with different pathogens, we focused on human as host.

The network inference step with *A. fumigatus* as fungal pathogen enriched the pathogenicity-associated genes significantly to a fraction of 24.4% (see **Figure 4A**). Further, the localization filtering for potential host–fungal interfaces also enriched the pathogenicity-relevant genes significantly to a fraction of 32.2%. At last, the refinement step for interactors associated to pathogenicity-relevant processes enriched the fraction to 39.5% (see **Figure 4A**). For human interactors with *C. albicans* as pathogen, we observed a similar enrichment of pathogenicity-associated genes from the protein-coding genes (16.5%) over the inferred (24.4%) and the localization-filtered (33.3%) to the pathogenicity-associated process refined (39.3%) interactors (see **Figure 4C**).

Due to the lack of knowledge about *C. albicans* and *A. fumigatus* PHIs, we exploited information of the databases PHI-base (Winnenburg et al., 2008) and PHIDIAS (Xiang et al., 2007) about experimentally validated virulence-associated genes. For the fungal pathogen *A. fumigatus,* we found 39 pathogenicity-associated genes in PHI-base and 29 genes in PHIDIAS (with an overlap of 14 genes), while for *C. albicans* 128 genes were found

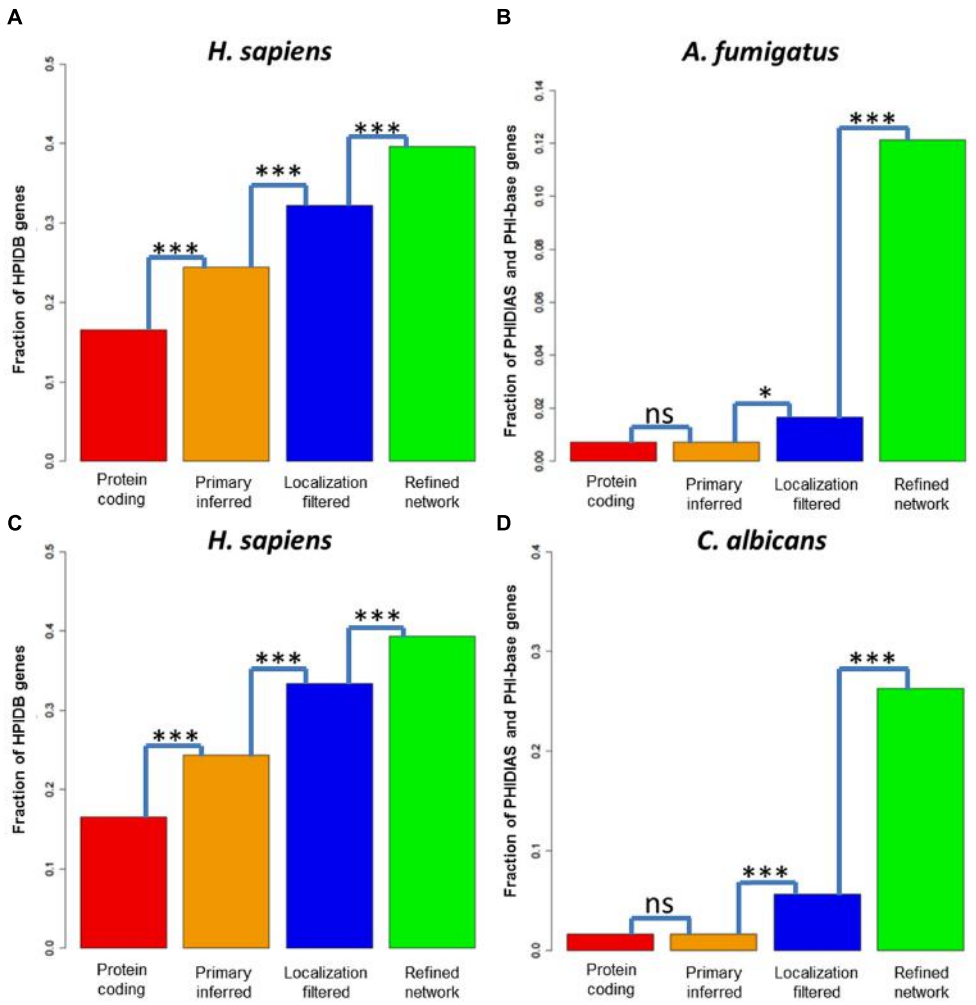


FIGURE 4 | Pathogenicity-associated genes in predicted host and fungi interactors. Fraction of pathogenicity-associated genes in the sets of protein-coding genes, the primary inferred, the localization-filtered and the biological process refined interactors of **(A)** *H. sapiens* **(B)** *Aspergillus fumigatus* **(C)** *H. sapiens* **(D)** *C. albicans*. A test for enrichment of pathogenicity-associated genes in the interactor sets after the primary inference, localization and functional refinement step (Fisher exact test,*$p < 0.05$; ***$p < 0.001$).

in PHI-base and 100 genes in PHIDIAS (with an overlap of 35 genes).

For the fungal pathogen *A. fumigatus*, the fraction of pathogenicity-relevant genes (0.7%) interacting with human genes was not significant for the interolog-based inference step (0.7%), weakly significant for the localization filtering step (1.7%) and strongly significant for the infection-relevant process refinement step (12.1%), (see **Figure 4B**). Similarly, the fraction of pathogenicity-associated genes (1.6%) did not increase significantly via the interolog-based inference step (1.6%), but strongly significant for the localization filtering step (5.6%) and strongly significant for the infection-relevant process refinement step (26.3%), (see **Figure 4D**).

Cells Involved in Immune Response and Tissues Typically Infected by Fungal Pathogens in the Resulting Host–Fungi PPI Networks

The tissue enrichment of refined *H. sapiens* interactors with either *C. albicans* or *A. fumigatus* and the primary *H. sapiens* interactors yielded several fungal infection relevant tissues (see Supplementary Tables S5 and S6). For both pathogens the cell type "Platelet" was most significantly enriched. This correlates with an investigation that attachment of platelets to fungal surfaces induced morphological changes in *Candida* spp., such as loosening of discoid shape, generation of pseudopodia, and flattened structure (Robert et al., 2000). Similar findings were described for *A. fumigatus* showing that hyphal growth is likely to induce platelet activation (Rodland et al., 2010). More in particular, certain cell wall components of *A. fumigatus*, e.g., melanin and galactosaminogalactan were involved in platelet activation while hydrophobin prevented recognition from the host immune system (Rambach et al., 2015). Besides platelets, the immune system-associated terms "B-cell lymphoma," "T-cell," "B-cell," "Leukemic T-cell," and "Peripheral blood lymphocyte" were significantly enriched. Furthermore, we observed significantly enriched tissue terms of typical environments of *Aspergillus* and *Candida* infections in the human body ("Lung," "Epithelium," "Blood," "Brain," and "Skin"). Interestingly, the tissues "Urinary bladder" and "Cervix" but also "Bone" were significantly enriched (see Supplementary Table S6).

Exploring the Refined Host–Fungi PPI Networks

To obtain an overview of the resulting refined networks, we visualized the interactors grouped by the functional GO slim biological process classes. Hence, the nodes represent GO slim terms and edges depict interactions between host and fungal genes belonging to the particular GO slim terms. Since the refined networks were dominated by few fungal interactors showing very high numbers of interactions, the top 10% of high degree fungal interactors (*C. albicans*: HSP90, UBI4, SSB1, SSA2, CaJ7_0234; *A. fumigatus*: glyceraldehyde-3-phosphate dehydrogenase GpdA, molecular chaperone and allergen Mod-E/Hsp90/Hsp1, 14-3-3 family protein ArtA) were removed from the network visualizations to improve clearness and readability of the figures (see **Figure 5** and Supplementary Figure S2).

In the *M. musculus* (330 interactors) and *C. albicans* (37 interactors) network, "signal transduction," "anatomical structure development," "cell differentiation," "response to stress," and "transport" represent the host GO slim terms consisting of the largest numbers of genes. For *Candida*, the terms comprising of the most interactors were "pathogenesis," "interspecies interaction between organisms," "filamentous growth," "response to stress," and "carbohydrate metabolic process." As expected, large murine GO slim terms frequently interact with large fungal GO slim terms (e.g., 795 interactions between "signal transduction" and "regulation of biological process" or 767 between "signal transduction" and "interspecies interaction between organisms"; see **Figure 5**).

In the refined PPI network with *H. sapiens* (317 interactors) and *A. fumigatus* (30 interactors), "signal transduction," "transport," "cellular nitrogen compound metabolic process," "response to stress," and "catabolic process" represent the host GO slim terms consisting of the largest numbers of genes. For *Aspergillus*, the terms comprising of the most interactors were "pathogenesis," "response to stress," "carbohydrate metabolic process," "response to chemical stimulus," and "cell cycle." Like for the mouse–*Candida* PPI network, large host GO slim terms frequently interact with large *Aspergillus* GO slim terms (e.g., 381 interactions between "signal transduction" and "pathogenesis" or 298 between "transport" and "pathogenesis"; see Supplementary Figure S2).

Mouse–*Candida* Subnetworks Contain Infection Related Interaction Candidates

To investigate these networks in more detail, we focused on the subnetwork between the pathogenicity-relevant GO slim terms "symbiosis, encompassing mutualism through parasitism" and "interspecies interaction between organisms" (see **Figure 6**). This subnetwork consists of 37 interactions with 23 murine interactors out of which one was infection regulated, and 12 *C. albicans* interactors of which three were infection regulated and eight supported by PHIDIAS/PHI-base evidence. For several interaction candidates, we found additional evidence in a literature research.

ENO1 and Cd4

One of those is the *Candida* ENO1 (2-phospho-D-glycerate-hydrolyase) interacting with the mouse Cd4 (CD4 antigen). The Cd4 molecule is an important co-receptor of T-lymphocytes that interacts with MHC Class II antigens. It is expressed in several immune cell types and initiates or augments the early phase of T-cell activation (Gibbings and Befus, 2009). The predicted interaction partner on the pathogen side, ENO1, is not only a key component of glycolysis (Sundstrom and Aliaga, 1992), but is also an immunodominant antigen circulating in the bloodstream of patients with disseminated *Candida* infections (Sundstrom and Aliaga, 1992) and a highly immunogenic protein in *Candida*-infected mice (Pitarch et al., 2001). Moreover, ENO1 was identified as an antigen that induced protective IgG2a antibody isotype in the sera from vaccinated animals and is thus considered a potential candidate for a vaccine (Fernandez-Arenas et al., 2004). Although ENO1 is primarily a cytoplasmic

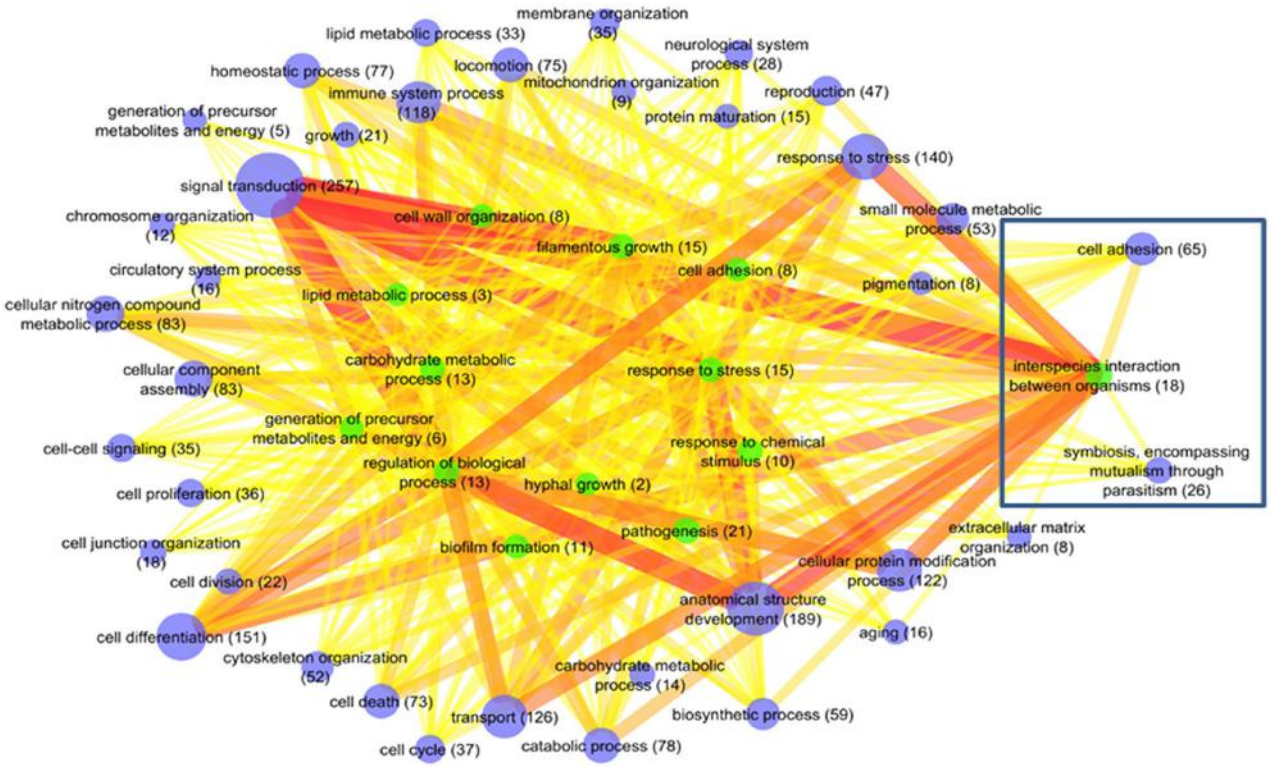


FIGURE 5 | *Mus musculus–C. albicans* network of functional GO terms. Nodes represent GO slim terms and edges depict interactions between host and fungal genes belonging to the particular GO slim terms. The node size denotes the number of genes in each GO slim term. The edge width and edge color correspond to the number of interactions between the connected nodes from thin/yellow to thick/red representing low to high interaction degrees. Fungal GO slim terms are visualized by green nodes and murine GO slim terms by blue nodes. The top 10% fungal pathogen interactors with the most interactions were removed from the network visualization to improve readability of the figure. The box shows the subnetworks that are evaluated in more detail.

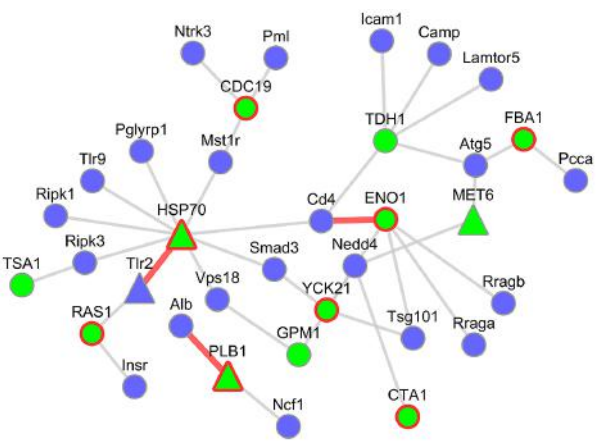


FIGURE 6 | Host–pathogen PPI subnetwork between *M. musculus* and *C. albicans*. This subnetwork comprises host interactors annotated as "symbiosis, encompassing mutualism through parasitism" and pathogen interactors annotated as "interspecies interaction between organisms." Blue nodes represent host interactors and green nodes fungal interactors. Nodes with a red border showed evidence for virulence contribution (PHIDIAS, PHI-base, and CGD). A triangular shape depicts infection-regulated genes of the analyzed mouse–*Candida* RNA-seq data. Interactions highlighted by red edges are described in more detail.

protein, it has also been discovered to be an integral cell wall protein (Angiolella et al., 1996). Interestingly, another infection-associated interaction partner in the refined PHI network is plasminogen, the inactive precursor of plasmin which has been described to facilitate the invasion of the host tissues (Jong et al., 2003).

PLB1 and Alb

A further interesting candidate is the interaction between the murine Alb (Albumin) and *Candida* PLB1 (Phospholipase B). It has been described that the extracellular part of PLB1 is required for wild-type virulence of *Candida* in a mouse model of systemic infection (Ghannoum, 1998), possibly related to its secretion from the hyphal tip during the infection process (Ghannoum, 2000). PLB1 can penetrate wild-type host cells by lysing the plasma membrane (Park et al., 2013). Its interaction partner on the host side, Albumin, was shown to bind to germ-tubes (Page and Odds, 1988) and to inhibit the binding of PLB1 to its substrate (Reisfeld et al., 1994). In the transcriptome data set of murine innate immune cells infected by *C. albicans*, PLB1 was significantly deregulated.

HSP70 and Tlr2

Heat shock proteins have been described to play a role during fungal infection (Lopez-Ribot et al., 1996). Our results predict an

interaction between the *Candida* HSP70 (Heat shock protein 70) and the murine Tlr2 (Toll-like receptor 2). The *Candida* HSP70 was detected on the surface of both yeast form and hyphal form cells (Urban et al., 2003) and is a member of a protein family which represents highly conserved immunodominant antigens (La Valle et al., 1995). *In vitro* studies showed that a *Candida* HSP70 mutant caused less damage to endothelial cells and oral epithelial cell lines (Sun et al., 2010). On the host side Tlr2 plays an important role in the activation of the innate immunity: It belongs to the family of pattern recognition receptors (PRRs) which are involved in the recognition of pathogen-associated molecular patterns (PAMPs), (Oliveira-Nascimento et al., 2012). Interestingly, the transcripts of both interaction partners were differentially upregulated during the infection process in the mouse–*Candida* dual RNA-seq experiment.

The mouse–*Candida* subnetwork of the host GO slim term "cell adhesion" and the fungal GO slim term "interspecies interaction between organisms" consisted of 98 interactions with 54 murine interactors (two significantly deregulated) and 16 *C. albicans* interacting partners (4 significantly deregulated, 11 supported by PHIDIAS/PHI-base evidence; see Supplementary Figure S3).

PLB1 and App

For the fungal PLB1 (Phospholipase B), we discovered a further potential interaction to the murine App [amyloid beta (A4) precursor protein]. APP is a cell surface receptor that mediates cell–cell and cell-matrix adhesion (Stahl et al., 2014) and is cleaved by secretases to form a number of peptides. Although, the human APP is primarily known for its role in Alzheimer's Disease (Gorevic et al., 1986), some of the App peptides have antibiotic activity against at least eight common and clinically relevant microorganisms, i.e., Gram-negative, Gram-positive bacteria, and the yeast *C. albicans* with the latter being the most sensitive (Soscia et al., 2010).

CDC19 and Egfr

We also found evidence for a very interesting interaction between the fungal CDC19 protein (Pyruvate kinase CDC19) and the murine Egfr protein (epidermal growth factor receptor). The fungal interactor CDC19, usually, an enzyme of the glycolysis, was found to be present on the yeast-form cell surface of *C. albicans* (Pitarch et al., 2002) and differentially expressed after 3-h co-culture with murine macrophages (Fernandez-Arenas et al., 2007). Furthermore, it is an immunogenic protein that is specifically recognized by antibodies in sera of vaccinated and of systemically *Candida*-infected mice (Pitarch et al., 2001; Thomas et al., 2006; Martinez-Lopez et al., 2008). A homozygous null mutant showed decreased virulence and filamentous growth (Binkley et al., 2014). Egfr is a transmembrane glycoprotein and receptor of the epidermal growth factor family. Egfr was shown to induce endocytosis of *C. albicans* by epithelial cells (Zhu et al., 2012). Furthermore, there is evidence for the secreted agrA (Accessory gene regulator protein A) of *Staphylococcus aureus* to bind to Egfr and activate a signal pathway in a pathogenicity-associated process (Gomez et al., 2007).

Examples for Interesting Human–*Aspergillus* PPIs in the Resulting Host–Fungi Network

Since very little is known about human–*Aspergillus* interactions in available databases up to date, we selected the infection-relevant subnetwork of interactions between the host GO slim term "symbiosis, encompassing mutualism through parasitism" and the fungal GO slim term "pathogenesis." To get a transparent

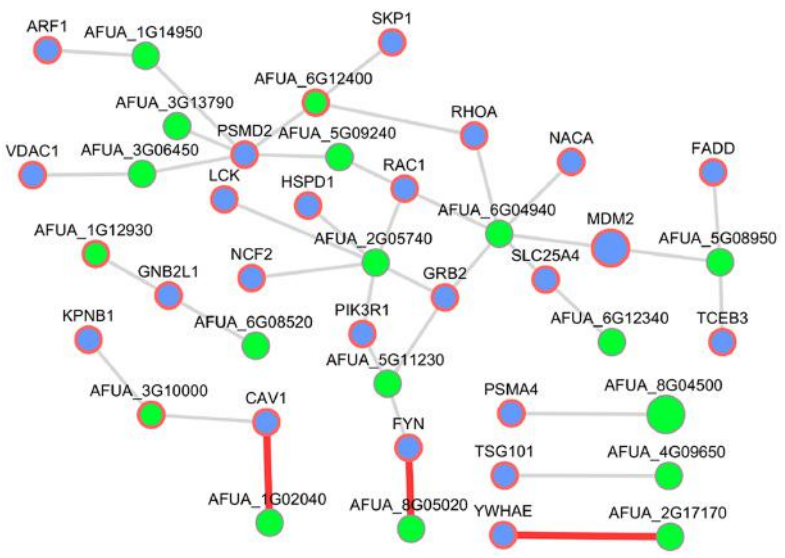


FIGURE 7 | Host–pathogen PPI subnetwork between *H. sapiens* and *A. fumigatus*. This subnetwork comprises pathogenicity-associated (HPIDB) host interactors annotated as "symbiosis, encompassing mutualism through parasitism" and pathogen interactors annotated as "pathogenesis." Blue nodes represent host interactors and green nodes fungal interactors. Nodes with a red border showed evidence for virulence contribution (PHIDIAS, PHI-base, and AspGD) or other host–pathogen interactions (HPIDB). Interactions highlighted by red edges were described in more detail.

size, we visualized only host nodes pathogenicity-associated based on HPIDB and removed the human interactor UBC (ubiquitin C) due to the high number of interactions. This subnetwork consists of 38 interactions with 23 human interactors and 18 *A. fumigatus* interacting partners (three supported by PHIDIAS/PHI-base evidence; see **Figure 7**).

RBE1 and CAV

The interesting interaction between the human CAV1 (caveolin 1) and the *Aspergillus* AFUA_1G02040 (Uncharacterized protein) in that subnetwork was inferred from the human template CAV1 – GLIPR2 (GLI pathogenesis-related 2) detected by affinity chromatography technology (Eberle et al., 2002). The *C. albicans* ortholog of AFUA_1G02040, RBE1 (Repressed by EFG1 protein 1), is a Pry family cell wall protein (Sohn et al., 2003) and belongs to a group of plant pathogenesis-related proteins (PR-1; Rohm et al., 2013). A homozygote null mutant of RBE1 in *Candida* showed a decreased virulence and increased sensitivity to attack by polymorphonuclear leucocytes (Rohm et al., 2013). The human CAV1 is the major structural protein in the caveolae of endothelial cells (Smart et al., 1999). It is also involved in the costimulatory signal essential for T-cell receptor (TCR)-mediated T-cell activation (Ohnuma et al., 2007) and can act as a functional receptor for CD26 in antigen representing cells (Ohnuma et al., 2004) which implies a cell surface localization.

CNH1 and YWHAE

In addition, we discovered another promising interaction, namely between the human YWHAE (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein) – AFUA_2G17170 (Uncharacterized protein) which is an ortholog of the fungal-specific *C. albicans* Na^+/H^+ antiporter CNH1 (Inglis et al., 2012). Homozygous null mutants of *Candida* ortholog showed decreased virulence (Soong et al., 2000). The human YWHAE, member of the 14-3-3 protein family was co-immunoprecipitated with MHC II in B-cell exosomes (Buschow et al., 2010) and thus implying an immune response relevant function.

HEX1 and FYN

In the human–*Aspergillus* subnetwork, we predicted an interaction between the human FYN (FYN Proto-oncogene) and the *Aspergillus* AFUA_8G05020 (Uncharacterized protein). FYN is a membrane-associated tyrosine kinase (Morford et al., 2002) and localized in the endosome (Puertollano, 2005). Further, it plays an important role in T-cell activation (Lancki et al., 1995). The *Aspergillus* AFUA_8G05020 is a putative secreted *N*-acetylhexosaminidase (Bruns et al., 2010; Sharma et al., 2011) which is highly expressed in biofilm (Bruns et al., 2010). Furthermore, the *C. albicans* ortholog HEX1 is required for full virulence and these proteins may have a role in carbon or nitrogen scavenging (Niimi et al., 1997).

Discussion

Even though fungal infections are clinically highly relevant and impose a substantial disease burden worldwide (Brown et al., 2012), not much data about interactions between fungal pathogens and the human host on a molecular level are currently available. In our study, a comprehensive search of publically available PHIs (Kumar and Nanduri, 2010) yielded only a small number of reported host–fungi PPIs. Also, thorough searches of all major PPI databases for cross-species interaction revealed only a few fungal candidates. This obvious sparseness of established experimental data on molecular host–fungal interactions generates an important and valuable research challenge for novel PHI prediction approaches. While *in silico* methods for the prediction of molecular interactions between host and pathogenic organisms have been receiving growing attention in the last years, the main focus still lays on viral and bacterial pathogens (Zhou et al., 2013a), and fungal species have only been sparsely investigated. To our knowledge, a thorough systematic prediction and analysis of *A. fumigatus* and *C. albicans* interactions with the human and murine host has not been performed so far.

In this study, we developed and examined an interolog-based method for the prediction of fungal–host interactions. We focused our investigation on two of the most clinically relevant fungi *C. albicans* and *A. fumigatus*. Since murine mouse models have become an invaluable tool in medical research, we also investigated interactions between these fungi and *M. musculus* in addition to the human host. As the primary objective of our study was to attain a comprehensive catalog of high quality PHI predictions, we used an extended dual species template approach which is based on human and yeast, the two best studied species for PPI network. By this we effectively made use of the majority of all publically available PPI data. Compared to simple approaches relying on the yeast template only, we created a considerably enhanced prediction space, in particular on the host side, which increases the set of interactors for human and mouse by over 200%.

A potential limitation of interspecies interolog approaches is the fact that the prediction space is confined to interactions between proteins with orthologs counterparts in the source network on either side. Hence, basing a prediction approach exclusively on the yeast network could lead to a bias toward ancient well conserved proteins and exclude less conserved 'newer' genes and pathways. These could include also host-specific genes such as those involved in novel adaptive immune responses. The inclusion of the human template network partially alleviates those effects as, at least on the host side, no basal orthology relationship is required. Our results suggest that a large and in particular human based template network is a key prerequisite for the prediction of functionally more relevant interactions.

Nevertheless, homology based approaches are known to be prone to produce overpredictions, since, in the first step, pairwise interactions are inferred between all homologs regardless of their cellular function or localization. Indeed, the predicted interaction partners on either side may in fact have little opportunity to physically interact with each other. This applies in particular to proteins which are expressed exclusively in the intracellular compartment and might thus have little opportunity to interact with the predicted host/pathogen

counterpart. Although we applied a rigorous filtering cascade to exclude many (99.4%) of these potentially spurious interaction predictions, we noted that many proteins are expressed in various subcellular compartments. In particular, numerous intracellular proteins can shuttle to the membrane compartment or even be secreted. To narrow down this set of 'potentially physically possible' predictions, we focused on interactors involved in pathways which play important roles during cellular infection processes.

Enrichment analyses using independent data (Xiang et al., 2007; Winnenburg et al., 2008; Kumar and Nanduri, 2010) revealed a clearly increasing fraction of virulence and pathogenicity-associated genes during the refinement process, suggesting a large set of functionally relevant interactions among the predictions. Moreover, on the host side we found an enrichment of genes which are expressed in tissues that are specifically affected by fungal infections, e.g., activation of platelets by *A. fumigatus* (Rodland et al., 2010) and *C. albicans* (Robert et al., 2000).

Our extended interolog-based approach assembled a large catalog of PHIs. As this homology based approach is tied to the template interaction network, it is confined to the set of reported physical PPIs and thus also inherits the set false positives from the template network. Therefore, an interesting complementary approach would be the investigation of an approach based on domain–domain interactions (Zhou et al., 2013b). This would eliminate the necessity of homology for the predicted interactors, as it only requires the presence of the interacting domains. Thus, it can be expected to yield a complementary dataset. Similarly, inference methods based on the correlated gene expression in host and pathogen (e.g., measured over an infection time course), are an interesting approach which could be further explored, in combination with and in comparison to the interolog approach (Wang et al., 2013; Weber et al., 2013; Schulze et al., 2015). Certainly, the assembly of large PHI networks establishes an ample hypotheses space as a basis which can be exploited by advanced methods of integrative network analysis (Dittrich et al., 2008; Beisser et al., 2012), for which a large number of approaches have been established in the last years. Here, further development is needed to extend these approaches to the simultaneous analysis of the complex connected host and pathogen networks. Albeit, technically not trivial, it is unquestionably a worthwhile task as it holds the potential to link subcellular response pathways between host and pathogen during the dynamics of the infection process.

Acknowledgment

The authors gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 "Pathogenic fungi and their human host: Networks of interaction," subproject B2.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00764

References

Angiolella, L., Facchin, M., Stringaro, A., Maras, B., Simonetti, N., and Cassone, A. (1996). Identification of a glucan-associated enolase as a main cell wall protein of *Candida albicans* and an indirect target of lipopeptide antimycotics. *J. Infect. Dis.* 173, 684–690. doi: 10.1093/infdis/173.3.684

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S., Ceol, A., Chautard, E., et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8, 528–529. doi: 10.1038/nmeth.1637

Arnaud, M. B., Costanzo, M. C., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., et al. (2005). The Candida Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.* 33, D358–D363. doi: 10.1093/nar/gki003

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND–the biomolecular interaction network database. *Nucleic Acids Res.* 29, 242–245. doi: 10.1093/nar/29.1.242

Beisser, D., Brunkhorst, S., Dandekar, T., Klau, G. W., Dittrich, M. T., and Muller, T. (2012). Robustness and accuracy of functional modules in integrated network analysis. *Bioinformatics* 28, 1887–1894. doi: 10.1093/bioinformatics/bts265

Binkley, J., Arnaud, M. B., Inglis, D. O., Skrzypek, M. S., Shah, P., Wymore, F., et al. (2014). The *Candida* Genome Database: the new homology information page highlights protein similarity and phylogeny. *Nucleic Acids Res.* 42, D711–D716. doi: 10.1093/nar/gkt1046

Bonfante, P., and Genre, A. (2010). Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. *Nat. Commun.* 1, 48. doi: 10.1038/ncomms1046

Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., et al. (2013). InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. *Nucleic Acids Res.* 41, D1228–D1233. doi: 10.1093/nar/gks1147

Brown, G. D., Denning, D. W., Gow, N. A., Levitz, S. M., Netea, M. G., and White, T. C. (2012). Hidden killers: human fungal infections. *Sci. Transl. Med.* 4, 165rv113. doi: 10.1126/scitranslmed.3004404

Brown, K. R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* 8, R95. doi: 10.1186/gb-2007-8-5-r95

Bruns, S., Seidler, M., Albrecht, D., Salvenmoser, S., Remme, N., Hertweck, C., et al. (2010). Functional genomic profiling of *Aspergillus fumigatus* biofilm reveals enhanced production of the mycotoxin gliotoxin. *Proteomics* 10, 3097–3107. doi: 10.1002/pmic.201000129

Buschow, S. I., Van Balkom, B. W., Aalberts, M., Heck, A. J., Wauben, M., and Stoorvogel, W. (2010). MHC class II-associated proteins in B-cell exosomes and potential functional implications for exosome biogenesis. *Immunol. Cell Biol.* 88, 851–856. doi: 10.1038/icb.2010.64

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* 10, 690–691. doi: 10.1038/nmeth.2561

Cerqueira, G. C., Arnaud, M. B., Inglis, D. O., Skrzypek, M. S., Binkley, G., Simison, M., et al. (2014). The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. *Nucleic Acids Res.* 42, D705–D710. doi: 10.1093/nar/gkt1029

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, D816–D823. doi: 10.1093/nar/gks1158

Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N., and Ricard-Blum, S. (2011). MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* 39, D235–D240. doi: 10.1093/nar/gkq830

Chen, Y. Y., Chao, C. C., Liu, F. C., Hsu, P. C., Chen, H. F., Peng, S. C., et al. (2013). Dynamic transcript profiling of *Candida albicans* infection in zebrafish: a pathogen-host interaction study. *PLoS ONE* 8:e72483. doi: 10.1371/journal.pone.0072483

Costanzo, M. C., Engel, S. R., Wong, E. D., Lloyd, P., Karra, K., Chan, E. T., et al. (2014). *Saccharomyces* genome database provides new regulation data. *Nucleic Acids Res.* 42, D717–D725. doi: 10.1093/nar/gkt1158

de Groot, P. W., Bader, O., De Boer, A. D., Weig, M., and Chauhan, N. (2013). Adhesins in human fungal pathogens: glue with plenty of stick. *Eukaryot. Cell* 12, 470–481. doi: 10.1128/EC.00364-12

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., and Muller, T. (2008). Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* 24, i223–i231. doi: 10.1093/bioinformatics/btn161

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Durmus Tekir, S., Cakir, T., Ardic, E., Sayilirbas, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen-host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Dyer, M. D., Murali, T. M., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 23, i159–i166. doi: 10.1093/bioinformatics/btm208

Eberle, H. B., Serrano, R. L., Fullekrug, J., Schlosser, A., Lehmann, W. D., Lottspeich, F., et al. (2002). Identification and characterization of a novel human plant pathogenesis-related protein that localizes to lipid-enriched microdomains in the Golgi complex. *J. Cell Sci.* 115, 827–838.

Fernandez-Arenas, E., Cabezon, V., Bermejo, C., Arroyo, J., Nombela, C., Diez-Orejas, R., et al. (2007). Integrated proteomics and genomics strategies bring new insight into *Candida albicans* response upon macrophage interaction. *Mol. Cell. Proteomics* 6, 460–478. doi: 10.1074/mcp.M600210-MCP200

Fernandez-Arenas, E., Molero, G., Nombela, C., Diez-Orejas, R., and Gil, C. (2004). Low virulent strains of *Candida albicans*: unravelling the antigens for a future vaccine. *Proteomics* 4, 3007–3020. doi: 10.1002/pmic.200400929

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* 42, D749–D755. doi: 10.1093/nar/gkt1196

Ghannoum, M. A. (1998). Extracellular phospholipases as universal virulence factor in pathogenic fungi. *Nippon Ishinkin Gakkai Zasshi* 39, 55–59. doi: 10.3314/jjmm.39.55

Ghannoum, M. A. (2000). Potential role of phospholipases in virulence and fungal pathogenesis. *Clin. Microbiol. Rev.* 13, 122–143. doi: 10.1128/CMR.13.1.122-143.2000

Gibbings, D., and Befus, A. D. (2009). CD4 and CD8: an inside-out coreceptor model for innate immune cells. *J. Leukoc. Biol.* 86, 251–259. doi: 10.1189/jlb.0109040

Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H., Lamb, B. T., and Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics* 24, 1743–1744. doi: 10.1093/bioinformatics/btn285

Gomez, M. I., Seaghdha, M. O., and Prince, A. S. (2007). *Staphylococcus aureus* protein A activates TACE through EGFR-dependent signaling. *EMBO J.* 26, 701–709. doi: 10.1038/sj.emboj.7601554

Gorevic, P. D., Goni, F., Ponsestel, B., Alvarez, F., Peress, N. S., and Frangione, B. (1986). Isolation and partial characterization of neurofibrillary tangles and amyloid plaque core in Alzheimer's disease: immunohistological studies. *J. Neuropathol. Exp. Neurol.* 45, 647–664. doi: 10.1097/00005072-198611000-00004

Gow, N. A., Van De Veerdonk, F. L., Brown, A. J., and Netea, M. G. (2012). *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. *Nat. Rev. Microbiol.* 10, 112–122.

Havlickova, B., Czaika, V. A., and Friedrich, M. (2008). Epidemiological trends in skin mycoses worldwide. *Mycoses* 51(Suppl. 4), 2–15. doi: 10.1111/j.1439-0507.2008.01606.x

Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* 9, 811–818. doi: 10.1002/sim.4780090710

Horn, F., Heinekamp, T., Kniemeyer, O., Pollmacher, J., Valiante, V., and Brakhage, A. A. (2012). Systems biology of fungal infection. *Front. Microbiol.* 3:108. doi: 10.3389/fmicb.2012.00108

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57.

Hube, B. (2004). From commensal to pathogen: stage- and tissue-specific gene expression of *Candida albicans*. *Curr. Opin. Microbiol.* 7, 336–341. doi: 10.1016/j.mib.2004.06.003

Inglis, D. O., Arnaud, M. B., Binkley, J., Shah, P., Skrzypek, M. S., Wymore, F., et al. (2012). The *Candida* genome database incorporates multiple *Candida* species: multispecies search and analysis tools with curated gene and protein information for *Candida albicans* and *Candida glabrata*. *Nucleic Acids Res.* 40, D667–D674. doi: 10.1093/nar/gkr945

Jong, A. Y., Chen, S. H., Stins, M. F., Kim, K. S., Tuan, T. L., and Huang, S. H. (2003). Binding of *Candida albicans* enolase to plasmin(ogen) results in enhanced invasion of human brain microvascular endothelial cells. *J. Med. Microbiol.* 52, 615–622. doi: 10.1099/jmm.0.05060-0

Krishnadev, O., and Srinivasan, N. (2011). Prediction of protein-protein interactions between human host and a pathogen and its application to three pathogenic bacteria. *Int. J. Biol. Macromol.* 48, 613–619. doi: 10.1016/j.ijbiomac.2011.01.030

Kumar, R., and Nanduri, B. (2010). HPIDB–a unified resource for host-pathogen interactions. *BMC Bioinformatics* 11(Suppl. 6):S16. doi: 10.1186/1471-2105-11-S6-S16

Lancki, D. W., Qian, D., Fields, P., Gajewski, T., and Fitch, F. W. (1995). Differential requirement for protein tyrosine kinase Fyn in the functional activation of antigen-specific T lymphocyte clones through the TCR or Thy-1. *J. Immunol.* 154, 4363–4370.

La Valle, R., Bromuro, C., Ranucci, L., Muller, H. M., Crisanti, A., and Cassone, A. (1995). Molecular cloning and expression of a 70-kilodalton heat shock protein of *Candida albicans*. *Infect. Immun.* 63, 4039–4045.

Lee, S. A., Chan, C. H., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., et al. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinformatics* 9(Suppl. 12):S11. doi: 10.1186/1471-2105-9-S12-S11

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., et al. (2012). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 40, D857–D861. doi: 10.1093/nar/gkr930

Lopez-Ribot, J. L., Alloush, H. M., Masten, B. J., and Chaffin, W. L. (1996). Evidence for presence in the cell wall of *Candida albicans* of a protein related to the hsp70 family. *Infect. Immun.* 64, 3333–3340.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Martin, F., and Nehls, U. (2009). Harnessing ectomycorrhizal genomics for ecological insights. *Curr. Opin. Plant Biol.* 12, 508–515. doi: 10.1016/j.pbi.2009.05.007

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *Bioinform. Action* 17, 10–12. doi: 10.14806/ej.17.1.200

Martinez-Lopez, R., Nombela, C., Diez-Orejas, R., Monteoliva, L., and Gil, C. (2008). Immunoproteomic analysis of the protective response obtained from vaccination with *Candida albicans* ecm33 cell wall mutant in mice. *Proteomics* 8, 2651–2664. doi: 10.1002/pmic.200701056

Morford, L. A., Forrest, K., Logan, B., Overstreet, L. K., Goebel, J., Brooks, W. H., et al. (2002). Calpain II colocalizes with detergent-insoluble rafts on human and Jurkat T-cells. *Biochem. Biophys. Res. Commun.* 295, 540–546. doi: 10.1016/S0006-291X(02)00676-9

Niimi, K., Niimi, M., Shepherd, M. G., and Cannon, R. D. (1997). Regulation of N-acetylglucosaminidase production in *Candida albicans*. *Arch. Microbiol.* 168, 464–472. doi: 10.1007/s002030050523

Ochman, H., and Moran, N. A. (2001). Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* 292, 1096–1098. doi: 10.1126/science.1058543

Ohnuma, K., Uchiyama, M., Yamochi, T., Nishibashi, K., Hosono, O., Takahashi, N., et al. (2007). Caveolin-1 triggers T-cell activation via CD26 in association with CARMA1. *J. Biol. Chem.* 282, 10117–10131. doi: 10.1074/jbc.M609157200

Ohnuma, K., Yamochi, T., Uchiyama, M., Nishibashi, K., Yoshikawa, N., Shimizu, N., et al. (2004). CD26 up-regulates expression of CD86 on antigen-presenting cells by means of caveolin-1. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14186–14191. doi: 10.1073/pnas.0405266101

Oliveira-Nascimento, L., Massari, P., and Wetzler, L. M. (2012). The role of TLR2 in infection and immunity. *Front. Immunol.* 3:79. doi: 10.3389/fimmu.2012.00079

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, D358–D363. doi: 10.1093/nar/gkt1115

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., et al. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* 9, 345–350. doi: 10.1038/nmeth.1931

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Page, S., and Odds, F. C. (1988). Binding of plasma proteins to *Candida* species in vitro. *J. Gen. Microbiol.* 134, 2693–2702. doi: 10.1099/00221287-134-10-2693

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., et al. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 832–834. doi: 10.1093/bioinformatics/bti115

Park, M., Do, E., and Jung, W. H. (2013). Lipolytic enzymes involved in the virulence of human pathogenic fungi. *Mycobiology* 41, 67–72. doi: 10.5941/MYCO.2013.41.2.67

Pitarch, A., Diez-Orejas, R., Molero, G., Pardo, M., Sanchez, M., Gil, C., et al. (2001). Analysis of the serologic response to systemic *Candida albicans* infection in a murine model. *Proteomics* 1, 550–559. doi: 10.1002/1615-9861(200104)1:4<550::AID-PROT550>3.0.CO;2-W

Pitarch, A., Sanchez, M., Nombela, C., and Gil, C. (2002). Sequential fractionation and two-dimensional gel analysis unravels the complexity of the dimorphic fungus *Candida albicans* cell wall proteome. *Mol. Cell. Proteomics* 1, 967–982. doi: 10.1074/mcp.M200062-MCP200

Prieto, C., and De Las Rivas, J. (2006). APID: agile protein interaction data analyzer. *Nucleic Acids Res.* 34, W298–W302. doi: 10.1093/nar/gkl128

Puertollano, R. (2005). Interactions of TOM1L1 with the multivesicular body sorting machinery. *J. Biol. Chem.* 280, 9258–9264. doi: 10.1074/jbc.M412481200

Rambach, G., Blum, G., Latge, J. P., Fontaine, T., Heinekamp, T., Hagleitner, M., et al. (2015). Identification of *Aspergillus fumigatus* surface components that mediate interaction of conidia and hyphae with human platelets. *J. Infect. Dis.* doi: 10.1093/infdis/jiv191 [Epub ahead of print].

Reisfeld, N., Lichtenberg, D., and Yedgar, S. (1994). Inhibition of LDL-associated phospholipase A activity in human plasma by albumin. *J. Basic Clin. Physiol. Pharmacol.* 5, 107–115. doi: 10.1515/JBCPP.1994.5.2.107

Rid, R., Strasser, W., Siegl, D., Frech, C., Kommenda, M., Kern, T., et al. (2013). PRIMOS: an integrated database of reassessed protein-protein interactions providing web-based access to in silico validation of experimentally derived data. *Assay Drug Dev. Technol.* 11, 333–346. doi: 10.1089/adt.2013.506

Robert, R., Nail, S., Marot-Leblond, A., Cottin, J., Miegeville, M., Quenouillere, S., et al. (2000). Adherence of platelets to *Candida* species in vivo. *Infect. Immun.* 68, 570–576. doi: 10.1128/IAI.68.2.570-576.2000

Rodland, E. K., Ueland, T., Pedersen, T. M., Halvorsen, B., Muller, F., Aukrust, P., et al. (2010). Activation of platelets by *Aspergillus fumigatus* and potential role of platelets in the immunopathogenesis of Aspergillosis. *Infect. Immun.* 78, 1269–1275. doi: 10.1128/IAI.01091-09

Rohm, M., Lindemann, E., Hiller, E., Ermert, D., Lemuth, K., Trkulja, D., et al. (2013). A family of secreted pathogenesis-related proteins in *Candida albicans*. *Mol. Microbiol.* 87, 132–151. doi: 10.1111/mmi.12087

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451. doi: 10.1093/nar/gkh086

Schulze, S., Henkel, S. G., Driesch, D., Guthke, R., and Linde, J. (2015). Computational prediction of molecular pathogen-host interactions based on dual transcriptome data. *Front. Microbiol.* 6:65. doi: 10.3389/fmicb.2015.00065

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, M., Soni, R., Nazir, A., Oberoi, H. S., and Chadha, B. S. (2011). Evaluation of glycosyl hydrolases in the secretome of *Aspergillus fumigatus* and saccharification of alkali-treated rice straw. *Appl. Biochem. Biotechnol.* 163, 577–591. doi: 10.1007/s12010-010-9064-3

Smart, E. J., Graf, G. A., Mcniven, M. A., Sessa, W. C., Engelman, J. A., Scherer, P. E., et al. (1999). Caveolins, liquid-ordered domains, and signal transduction. *Mol. Cell. Biol.* 19, 7289–7304.

Sohn, K., Urban, C., Brunner, H., and Rupp, S. (2003). EFG1 is a major regulator of cell wall dynamics in *Candida albicans* as revealed by DNA microarrays. *Mol. Microbiol.* 47, 89–102. doi: 10.1046/j.1365-2958.2003.03300.x

Sonnhammer, E. L., and Ostlund, G. (2014). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 43, D234–D239. doi: 10.1093/nar/gku1203

Soong, T. W., Yong, T. F., Ramanan, N., and Wang, Y. (2000). The *Candida albicans* antiporter gene CNH1 has a role in Na+ and H+ transport, salt tolerance, and morphogenesis. *Microbiology* 146(Pt 5), 1035–1044.

Soscia, S. J., Kirby, J. E., Washicosky, K. J., Tucker, S. M., Ingelsson, M., Hyman, B., et al. (2010). The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide. *PLoS ONE* 5:e9505. doi: 10.1371/journal.pone.0009505

Stahl, R., Schilling, S., Soba, P., Rupp, C., Hartmann, T., Wagner, K., et al. (2014). Shedding of APP limits its synaptogenic activity and cell adhesion properties. *Front. Cell. Neurosci.* 8:410. doi: 10.3389/fncel.2014.00410

Sun, J. N., Solis, N. V., Phan, Q. T., Bajwa, J. S., Kashleva, H., Thompson, A., et al. (2010). Host cell invasion and virulence mediated by *Candida albicans* Ssa1. *PLoS Pathog.* 6:e1001181. doi: 10.1371/journal.ppat.1001181

Sundstrom, P., and Aliaga, G. R. (1992). Molecular cloning of cDNA and analysis of protein secondary structure of *Candida albicans* enolase, an abundant, immunodominant glycolytic enzyme. *J. Bacteriol.* 174, 6789–6799.

The Gene Ontology, C. (2014). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056.

The UniProt Consortium. (2014). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212.

Thomas, D. P., Viudes, A., Monteagudo, C., Lazzell, A. L., Saville, S. P., and Lopez-Ribot, J. L. (2006). A proteomic-based approach for the identification of *Candida albicans* protein components present in a subunit vaccine that protects against disseminated candidiasis. *Proteomics* 6, 6033–6041. doi: 10.1002/pmic.200600321

Tierney, L., Linde, J., Muller, S., Brunke, S., Molina, J. C., Hube, B., et al. (2012). An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front. Microbiol.* 3:85. doi: 10.3389/fmicb.2012.00085

Tyagi, N., Krishnadev, O., and Srinivasan, N. (2009). Prediction of protein-protein interactions between *Helicobacter pylori* and a human host. *Mol. Biosyst.* 5, 1630–1635. doi: 10.1039/b906543c

Urban, C., Sohn, K., Lottspeich, F., Brunner, H., and Rupp, S. (2003). Identification of cell surface determinants in *Candida albicans* reveals Tsa1p, a protein differentially localized in the cell. *FEBS Lett.* 544, 228–235. doi: 10.1016/S0014-5793(03)00455-1

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., et al. (2000). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122. doi: 10.1126/science.287.5450.116

Wang, Y. C., Lin, C., Chuang, M. T., Hsieh, W. P., Lan, C. Y., Chuang, Y. J., et al. (2013). Interspecies protein-protein interaction network construction for characterization of host-pathogen interactions: a *Candida albicans*-zebrafish interaction study. *BMC Syst. Biol.* 7:79. doi: 10.1186/1752-0509-7-79

Weber, M., Henkel, S. G., Vlaic, S., Guthke, R., Van Zoelen, E. J., and Driesch, D. (2013). Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator V2.0. *BMC Syst. Biol.* 7:1. doi: 10.1186/1752-0509-7-1

Winnenburg, R., Urban, M., Beacham, A., Baldwin, T. K., Holland, S., Lindeberg, M., et al. (2008). PHI-base update: additions to the pathogen host interaction database. *Nucleic Acids Res.* 36, D572–D576. doi: 10.1093/nar/gkm858

Wuchty, S. (2011). Computational prediction of host-parasite protein interactions between *P. falciparum* and *H. sapiens*. *PLoS ONE* 6:e26960. doi: 10.1371/journal.pone.0026960

Xiang, Z., Tian, Y., and He, Y. (2007). PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol.* 8, R150. doi: 10.1186/gb-2007-8-7-r150

Zhou, H., Gao, S., Nguyen, N. N., Fan, M., Jin, J., Liu, B., et al. (2014). Stringent homology-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *Biol. Direct* 9:5. doi: 10.1186/1745-6150-9-5

Zhou, H., Jin, J., and Wong, L. (2013a). Progress in computational studies of host-pathogen interactions. *J. Bioinform. Comput. Biol.* 11, 1230001. doi: 10.1142/S0219720012300018

Zhou, H., Rezaei, J., Hugo, W., Gao, S., Jin, J., Fan, M., et al. (2013b). Stringent DDI-based prediction of *H. sapiens-M. tuberculosis* H37Rv protein-protein interactions. *BMC Syst. Biol.* 7(Suppl. 6):S6. doi: 10.1186/1752-0509-7-S6-S6

Zhu, W., Phan, Q. T., Boontheung, P., Solis, N. V., Loo, J. A., and Filler, S. G. (2012). EGFR and HER2 receptor kinase signaling mediate epithelial cell invasion by *Candida albicans* during oropharyngeal infection. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14194–14199. doi: 10.1073/pnas.1117676109

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 December 2015
doi: 10.3389/fmicb.2015.01386

Literature Mining and Ontology based Analysis of Host-*Brucella* Gene–Gene Interaction Network

İlknur Karadeniz[1], Junguk Hur[2], Yongqun He[3,4,5]* and Arzucan Özgür[1]**

[1] Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey, [2] Department of Basic Sciences, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, ND, USA, [3] Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA, [4] Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA, [5] Comprehensive Cancer Center, University of Michigan Health System, Ann Arbor, MI, USA

Brucella is an intracellular bacterium that causes chronic brucellosis in humans and various mammals. The identification of host-*Brucella* interaction is crucial to understand host immunity against *Brucella* infection and *Brucella* pathogenesis against host immune responses. Most of the information about the inter-species interactions between host and *Brucella* genes is only available in the text of the scientific publications. Many text-mining systems for extracting gene and protein interactions have been proposed. However, only a few of them have been designed by considering the peculiarities of host–pathogen interactions. In this paper, we used a text mining approach for extracting host-*Brucella* gene–gene interactions from the abstracts of articles in PubMed. The gene–gene interactions here represent the interactions between genes and/or gene products (e.g., proteins). The SciMiner tool, originally designed for detecting mammalian gene/protein names in text, was extended to identify host and *Brucella* gene/protein names in the abstracts. Next, sentence-level and abstract-level co-occurrence based approaches, as well as sentence-level machine learning based methods, originally designed for extracting intra-species gene interactions, were utilized to extract the interactions among the identified host and *Brucella* genes. The extracted interactions were manually evaluated. A total of 46 host-*Brucella* gene interactions were identified and represented as an interaction network. Twenty four of these interactions were identified from sentence-level processing. Twenty two additional interactions were identified when abstract-level processing was performed. The Interaction Network Ontology (INO) was used to represent the identified interaction types at a hierarchical ontology structure. Ontological modeling of specific gene–gene interactions demonstrates that host–pathogen gene–gene interactions occur at experimental conditions which can be ontologically represented. Our results show that the introduced literature mining and ontology-based modeling approach are effective in retrieving and analyzing host–pathogen gene–gene interaction networks.

Keywords: host–pathogen interaction extraction, *Brucella*, text mining, host and pathogen gene name recognition, SciMiner, support vector machines (SVM), Interaction Network Ontology (INO)

OPEN ACCESS

Edited by:
Awdhesh Kalia,
University of Texas MD Anderson Cancer Center, USA

Reviewed by:
Li Xu,
Cornell University, USA
Hao-Teng Chang,
China Medical University, Taiwan

****Correspondence:***
Arzucan Özgür
arzucan.ozgur@boun.edu.tr;
Yongqun He
yongqunh@med.umich.edu;
Junguk Hur
junguk.hur@med.und.edu

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *19 May 2015*
Accepted: *20 November 2015*
Published: *09 December 2015*

Citation:
Karadeniz İ, Hur J, He Y and Özgür A (2015) Literature Mining and Ontology based Analysis of Host-Brucella Gene–Gene Interaction Network. Front. Microbiol. 6:1386. doi: 10.3389/fmicb.2015.01386

INTRODUCTION

Brucella is a Gram-negative intracellular bacterium that causes zoonotic brucellosis in humans and various animals. Brucellosis is one of the most common zoonotic diseases worldwide, causing approximately half a million new human brucellosis each year. There are 10 species of *Brucella* based on the preferential host specificity: *Brucella melitensis* (goats), *B. abortus* (cattle), *B. suis* (swine), *B. canis (*dogs), *B. ovis* (sheep), *B. neotomae* (desert mice), *B. cetaceae* (cetacean), *B. pinnipediae* (seal), *B. microti* (voles), and *B. inopinata* (unknown) (O'Callaghan and Whatmore, 2011). Among them, *B. melitensis*, *B. abortus*, *B. suis,* and *B. canis* are pathogenic to human. The other *Brucella* species are non-pathogenic to humans.

The genome sequences of all *Brucella* species are strikingly similar with nearly identical genetic content and gene organization (Halling et al., 2005). Humans can be infected with *Brucella* by contact with infected animals, by inhalation of an aerosol, or by ingestion of contaminated animal products (e.g., infected milk and meat). Upon entry into animals, the bacteria invade the blood stream and lymphatics where they multiply inside phagocytic cells and eventually cause septicemia. Symptoms include undulant fever, abortion, asthenia, endocarditis and encephalitis. In spite of a long documented history (Corbel, 1997), the treatment of human brucellosis remains difficult and requires antibiotics that penetrate macrophages and can act in an acidic intracellular environment. While currently used live attenuated *Brucella* animal vaccines (e.g., RB51, strain 19, and Rev. 1) have the ability to protect animals, they are still pathogenic to humans. No safe and effective *Brucella* vaccine is available for human use. To develop safe and effective preventive and therapeutic measures against *Brucella* infections, it is critical to understand the host-*Brucella* mechanisms that lead to *Brucella* pathogenesis and host immunity against *Brucella* infection. Although extensive studies have been undertaken, the systematic understanding of the host-*Brucella* interactions is still missing.

Currently, there is very limited information regarding host-*Brucella* interactions in the host–pathogen interaction databases such as PHIDIAS (Xiang et al., 2007), PHISTO (Tekir et al., 2013), and HPIDB (Kumar and Nanduri, 2010). Most of the relevant information is only available in a textual format in the published scientific articles. In this study, our goal is to utilize text mining methods to extract host-*Brucella* gene interactions from the biomedical literature. In order to extract host–pathogen gene interactions, first the pathogen and host gene names should be identified in text, then the interactions among the host and pathogen genes should be detected. For example, the sentence shown in **Figure 1** (Arenas-Gamboa et al., 2008) contains three host genes (*gamma interferon, interleukin-12,* and *interleukin-4*) and one pathogen gene (*vjbR)*. This sentence states that there are two pathogen–host gene interactions: *(gamma interferon, vjbR)* and *(interleukin-12, vjbR).* On the other hand, there is no an interaction between the host gene *interleukin-4* and pathogen gene *vjbR*.

Different methods have been proposed for literature mining of gene–gene interactions. One of the simplest and widely used methods is based on the co-occurrence statistics of the proteins in text (Jelier et al., 2005). Another common approach is matching pre-specified patterns and rules over the sequences of words and/or their parts of speech in the sentences (Ono et al., 2001; Blaschke and Valencia, 2002). More recently, machine learning methods that integrate the linguistic, syntactic, and/or semantic analysis of the sentences as kernel functions have been proposed and shown to achieve state-of-the-art results for gene/protein interaction extraction from text (Giuliano et al., 2006; Erkan et al., 2007; Airola et al., 2008; Tikk et al., 2010). Similarly to previous literature mining studies, in this paper we used the commonly applied GENETAG-style named entity annotation (Tanabe et al., 2005). In other words, a gene interaction can involve genes or gene products such as proteins.

A number of rule-based and machine learning based methods have been proposed for identifying gene/protein mentions in text (Fukuda et al., 1998; McDonald and Pereira, 2005; Tsai et al., 2006; Hsu et al., 2008). In our previous studies, we developed dictionary- and rule-based named entity recognition tools, SciMiner (Hur et al., 2009) and Vaccine Ontology (VO)-SciMiner (Hur et al., 2011), which are designed to identify genes/proteins and Vaccine Ontology (VO) terms in the biomedical literature. Conventional Medical Subject Headings (MeSH) terminology has been frequently used for literature mining, such as GenoMesh studies (Xiang et al., 2013). The usage of ontologies enhances the chances of retrieving gene–gene interactions. For example, in our recent studies we have shown that the VO facilitates the retrieval of vaccine-associated IFN-gamma interaction network (Özgür et al., 2011), fever-related network (Hur et al., 2012), and *Brucella* vaccine interaction network (Hur et al., 2012). Recently, we have developed an Interaction Network Ontology (INO) which is used to classify the interaction keywords such as up-regulation, inhibition, association, and binding in an ontology structure (Hur et al., 2015). The classified interaction hierarchy makes us not only retrieve gene–gene interactions, but also the types of gene–gene interactions (Hur et al., 2015). We hypothesize that such a strategy can also be used in host–pathogen gene–gene interaction literature retrieval.

Currently, the research in host–pathogen interactions literature mining mostly focuses on the retrieval of host gene–gene interaction under a particular pathogen infection (e.g., influenza) or pathogen gene–gene interactions [e.g., our *Brucella* vaccine interaction network analysis (Hur et al., 2012)]. There are only a few studies on the retrieval of both host and pathogen genes and the inter-species interactions among them [reviewed in (Durmus et al., 2015)]. Machine learning based methods were proposed for classifying abstracts of scientific articles as being relevant to host–pathogen interactions or not (Yin et al., 2010; Thieu et al., 2012). In addition, Thieu et al. (2012) proposed a rule-based approach that is based on the link-grammar representations of the sentences for extracting host–pathogen protein interactions from text.

In this study, we use kernel-based methods for extracting host–pathogen gene interactions, which have been shown to achieve promising results for extracting intra-species protein interactions (Erkan et al., 2007; Tikk et al., 2010). One main issue in host–pathogen interaction literature mining is the confusion

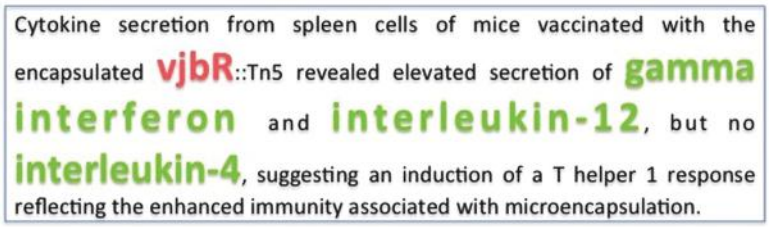


FIGURE 1 | Sample host–pathogen interaction describing sentence (Arenas-Gamboa et al., 2008). The pathogen gene is shown in red and the host genes are shown in green.

of a gene being a host gene or pathogen gene, since many gene names are shared in both hosts and pathogens. This is one main research topic in our current study. We extended the SciMiner mammalian gene name identification tool to recognize and distinguish between host and *Brucella* genes. In addition, we used an INO-based method to model various gene–gene interactions under different experimental conditions. Our results show that our combinatory strategy is able to successfully retrieve and analyze host–pathogen gene–gene interaction networks.

MATERIALS AND METHODS

The main focus of this study is to identify the interactions between host and *Brucella* genes. Many eukaryotic organisms act as the host of *Brucella* infections, including human, cattle, goat, sheep, pig, etc. As a laboratory animal model, mice can also be infected with *Brucella*. Our literature mining study covers these different host species. Meanwhile, there are 10 different *Brucella* species.

The overall design and workflow of our approach is shown in **Figure 2**. All PubMed papers are used as our data sources. They are filtered based on their relevance to *Brucella*. The selected abstracts are processed by splitting into sentences and identifying the host and *Brucella* gene name mentions using SciMiner. Next, co-occurrence and machine learning based methods are used to extract the interactions among the host and *Brucella* genes. A literature-mined and manually verified host-*Brucella* gene–gene interaction network is created. Finally, ontology based modeling of host–pathogen gene–gene interactions is performed by utilizing the INO. The details of the methods are presented in the following subsections.

Data Set Collection

The 2015 MEDLINE®/PubMed® Baseline Distribution database consisting of 23,343,329 records was downloaded from the US National Library of Medicine and processed using our established literature mining pipeline. Briefly, the title, abstract, and MeSH terms of each record were parsed out from the downloaded XML files. The collected abstracts were split into sentence level using Java's LBJ2.nlp.SentenceSplitter module. Then, enhanced version of our named entity recognition tools, SciMiner (Hur et al., 2009) and VO-SciMiner (Hur et al., 2011), were used to identify host genes and pathogen genes, and the results were populated into a local MySQL database. To define the *Brucella*-specific context, we used a PubMed query, "*Brucella* OR Brucellosis," which resulted in a list of 16,699 PubMed IDs as of 2/1/2015.

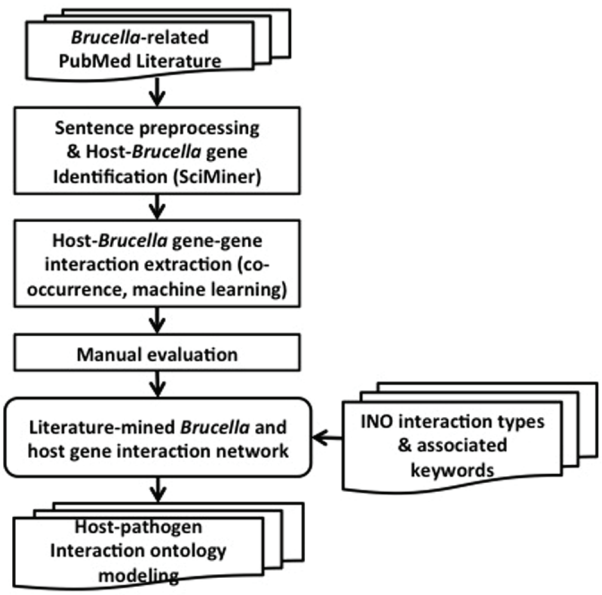


FIGURE 2 | Project design pipeline and workflow.

Identifying Gene Names

To identify the mentioned host genes and *Brucella* genes in the abstracts of articles, we used our in-house named entity recognizers, SciMiner[1] (Hur et al., 2009) and VO-SciMiner[2] (Hur et al., 2011). SciMiner and VO-SciMiner are both dictionary- and rule-based literature mining tools. SciMiner focuses on identification of mammalian genes, reported in terms of the official human genes based on the HUGO Gene Nomenclature Committee (HGNC) database[3], while VO-SciMiner identifies VO terms and *Brucella* genes.

In the present study, to improve identification accuracy of host and pathogen genes, we enhanced the mining rules in

[1]http://jdrf.neurology.med.umich.edu/SciMiner/
[2]http://www.violinet.org/vo-sciminer/
[3]http://www.genenames.org/

both SciMiner and VO-SciMiner. First, the enhanced version of SciMiner uses a stringent case-match of gene symbols. In the original version of SciMiner, which included dictionary of only human genes names and symbols, a relaxed matching of symbols was employed to maximize the gene identification (high recall). This relaxed case matching resulted in misidentifications such as *recA*, recombinase A gene, being identified as the human RAD51 recombinase (RAD51), whose aliases include RECA. Since the majority of the *Brucella* gene symbols start with a lower-case character and usually end with an upper-case or numeric character, SciMiner excluded symbols with this pattern. In case of the genes identified by both SciMiner as a host gene and VO-SciMiner as a pathogen gene, the priority is given to the VO-SciMiner identification considering the current context of *Brucella*-related literature.

Mapping Genes to Pathogen and Host Species

In order to further improve the overall accuracy of host gene identification, we used potential host species-related MeSH terms, including 'humans,' 'rats,' 'mice,' 'cattle,' 'guinea pigs,' 'swine,' 'goats,' and 'sheep' to filter the genes identified by SciMiner. Only the host genes identified from PubMed documents whose MeSH terms included at least one of these selected terms were included for further analysis.

Gene–gene Interaction Extraction

In this study, co-occurrence based and machine-learning based approaches are used for extracting host–pathogen gene–gene interactions. Both sentence-level and abstract-level co-occurrence approaches, as well as a machine learning-based approach are investigated for this task. These approaches are described in the following subsections.

Co-occurrence Based Host–pathogen Interaction Extraction

We used two different contexts to extract the interactions based on the co-occurrences of the host and pathogen genes: sentence-based context and abstract-based context. In the sentence-based co-occurrence approach, if one pathogen and one host gene occur in the same sentence, an interaction pair is extracted consisting of the corresponding pathogen and host genes. For example, in the sentence shown in **Figure 1** (Arenas-Gamboa et al., 2008), the SciMiner tool identifies two host genes *(interleukin-12* and *interleukin-4*) and one pathogen gene (*vjbR)*. The sentence-level co-occurrence approach extracts the interactions *(interleukin-12, vjbR)* and *(interleukin-4, vjbR)* from the sample sentence, where *(interleukin-12, vjbR)* is a true interaction and *(interleukin-4, vjbR)* is an incorrectly extracted interaction. In the sample sentence, *gamma interferon* is also a host gene. However, since this gene is not detected by SciMiner, it is not considered in the interaction extraction step. In the abstract-based co-occurrence approach, an abstract is taken into consideration as the context window instead of a single sentence. In other words, all pairs of host and pathogen genes that occur in the same abstract are extracted as interacting pairs regardless of the sentence boundaries.

Machine Learning Based Host–pathogen Interaction Extraction

We utilized a machine learning based approach to classify whether a host and pathogen gene pair occurring in the same sentence is described as interacting in the sentence or not. We used support vector machines (SVM) [specifically the SVM^{light} package (Joachims, 1999)] as our classification algorithm with the cosine and edit kernels introduced in (Erkan et al., 2007). These kernels make use of the dependency parse trees of the sentences that represent the syntactic and semantic relations among the words. We used the Stanford Parser (de Marneffe et al., 2006) to obtain the dependency parse trees of the sentences in our *Brucella* specific data set. We only processed sentences for which SciMiner identified at least one host and one pathogen gene. The cosine and edit kernels are defined over the path between the host gene and pathogen gene in the dependency parse tree of the corresponding sentence.

The underlying assumption is that the dependency path between a host and a pathogen gene is a good description for the relation between them. For example, the dependency parse tree obtained using the Stanford parser (de Marneffe et al., 2006) for the sample sentence "Furthermore, *gap* associated with murine *IL-12* gene in a DNA vaccine formulation partially protected mice against experimental infection." (Rosinha et al., 2002), is shown in **Figure 3**. The dependency path between the host gene *IL-12* and the pathogen gene *gap*, which are described as interacting in the given sentence, is "nn gene prep_with associated vmod." On this path we have the word *associated* as well as the dependency relation type *preposition with (prep_with)*, which provide clues for the interaction between *gap* and *IL-12*. Using the cosine similarity and edit distance kernel functions within SVM (Erkan et al., 2007), our program is able to infer whether or not these two genes interact with each other. Note that this sentence also includes the gene symbol "gap" which is a common English word. SciMiner has a confidence scoring system for each identified gene symbol in the text, based on weighted co-occurrences of the gene symbol and their descriptions (e.g., gene or protein names) in the same text. In this case, since the protein name of the gap gene "glyceraldehyde-3-phosphate dehydrogenase" is described in the paper abstract, the SciMiner scoring system was able to assign gap as a gene.

To the best of our knowledge, there are no publicly available manually labeled host–pathogen gene–gene interaction corpora. Therefore, we trained the SVM classifier with edit and cosine kernels by using corpora labeled for intra-species protein–protein interactions. Specifically, we used the Christina Brun (CB) corpus provided as a resource at the BioCreAtIve II challenge[4] and the AIMED corpus (Bunescu et al., 2005), which is a standard corpus for evaluating intra-species protein–protein interactions. The learned cosine and edit kernel based SVM models are used to classify each sentence as an interaction-describing sentence (positive class) or not (negative class) for each host and pathogen gene pair identified by SciMiner in the corresponding sentence.

[4]http://biocreative.sourceforge.net/biocreative_2.html

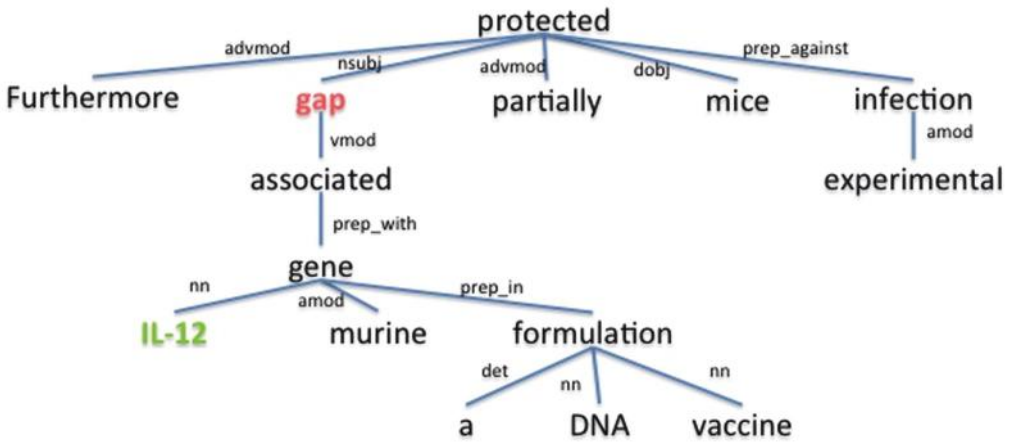


FIGURE 3 | The dependency parse tree of a sample sentence. The tree is generated for the sentence "Furthermore, gap associated with murine *IL-12* gene in a DNA vaccine formulation partially protected mice against experimental infection." from the abstract of (Rosinha et al., 2002). Host and pathogen genes identified by SciMiner are shown in green and red, respectively. The Stanford parser was used to generate the parse tree. advmod, adverb modifier; amod, adjectival modifier; det, determiner; dobj, direct object; nn, noun compound modifier; nsubj, nominal subject; prep_against, preposition against; prep_in, preposition in; prep_with, preposition with; vmod, reduced non-finite verbal modifier.

Evaluation

The results obtained by the co-occurrence and machine learning based interaction classification methods (i.e., classifiers) are manually evaluated by using the number of TP (True Positives), FP (False Positives), TN (True Negatives), and FN (False Negatives), as well as the precision, recall, and *F*-score metrics.

True Positives is the number of host–pathogen interactions correctly classified as positive; FP (False Positives) is the number of negative host–pathogen interactions that are incorrectly classified as positive by the classifier; TN (True Negatives) is the number of host pathogen interactions classified correctly as negative (no interaction); and FN (False Negatives) is the number of positive host–pathogen interactions that are incorrectly classified as negative by the classifier.

Precision is the ratio of correctly identified positive host–pathogen interactions over all interactions classified as positive by the classifier [i.e., TP/(TP + FP)]. Recall is the ratio of correctly classified positive host–pathogen interactions over all positive host–pathogen interactions [i.e., TP/(TP + FN)]. *F*-score is the harmonic mean of these two measures [i.e., 2 . precision . recall/(precision + recall)].

Ontology Modeling

The INO focuses on the ontological representations of hierarchical biological interaction types and networks (Hur et al., 2015). INO has been proven to enhance the literature mining of gene–gene interaction types (Hur et al., 2015). In this study, we applied INO to analyze different interaction types between host and *Brucella* at different experimental conditions. Furthermore, different conditions of host-*Brucella* interactions were represented and analyzed through ontology-based modeling.

RESULTS

Identification of Host and *Brucella* Gene Names

Two of our in-house named entity recognizers, SciMiner and VO-SciMiner, were enhanced in our study to identify host and pathogen genes, respectively. First, SciMiner has been modified to use stringent case-match. In the context of *Brucella*, consisting of 16,699 PubMed abstracts, the enhanced versions of SciMiner and VO-SciMiner identified 47 unique pairs of potential host gene and *Brucella* gene interactions using the improved symbol-based identification method and confliction resolution between host and *Brucella* gene. Out of these 47 pairs, manual examination confirmed that 24 unique pairs were true interactions, indicating an overall accuracy of 51%.

Identification of Host-*Brucella* Gene–gene Interactions

After identifying the host and *Brucella* gene names in sentences co-occurrence and machine learning based methods are used to classify each pair in a sentence as an interaction (positive class) or not (negative class). We performed manual evaluation for the classification decisions of the methods for each host-*Brucella* gene pair in each sentence. For the abstract-level co-occurrence approach, manual evaluation is performed for each host-*Brucella* gene pair in each abstract.

The results obtained are summarized in **Table 1**. Co-occurrence based methods classify all pairs of host–pathogen genes as positive, if they occur in the same sentence or abstract. Therefore, they obtain the maximum level of recall, i.e., 100%. Not all co-occurring gene pairs are true interaction pairs. For example, in the sample sentence shown in **Figure 1**, there is no an interaction between the pathogen gene *vjbR*

TABLE 1 | Co-occurrence and machine learning based host-*Brucella* gene–gene interaction results.

	TP	TN	FP	FN	Precision	Recall	*F*-score
Co-occurrence (sentence-based)	29	0	25	0	0.54	1.0	0.70
Co-occurrence (abstract-based)	55	0	61	0	0.47	1.0	0.64
Support vector machines (SVM; edit kernel)	15	12	12	14	0.56	0.52	0.54
SVM (cosine kernel)	12	19	5	17	0.71	0.41	0.52

TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative.

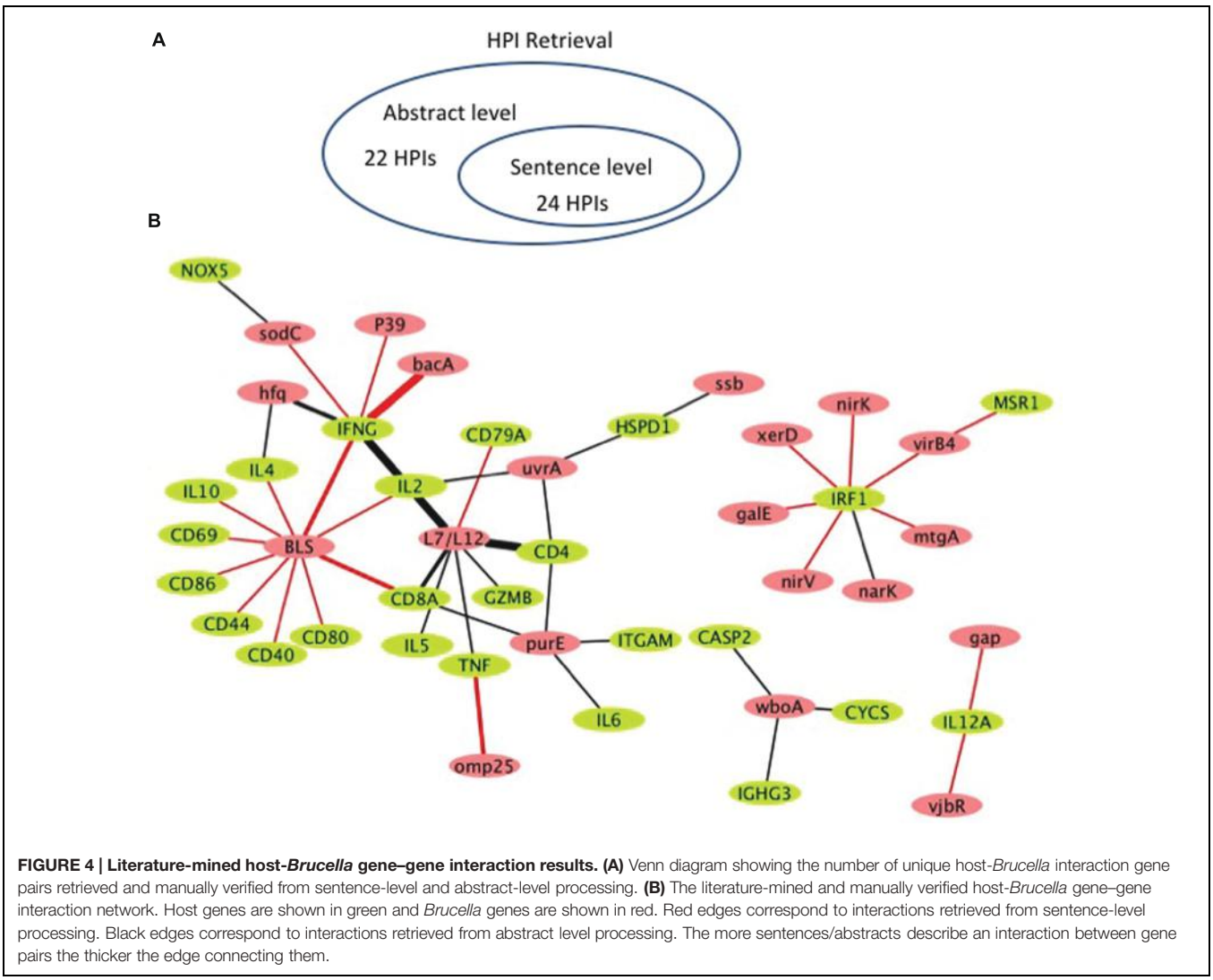


FIGURE 4 | Literature-mined host-*Brucella* gene–gene interaction results. (A) Venn diagram showing the number of unique host-*Brucella* interaction gene pairs retrieved and manually verified from sentence-level and abstract-level processing. **(B)** The literature-mined and manually verified host-*Brucella* gene–gene interaction network. Host genes are shown in green and *Brucella* genes are shown in red. Red edges correspond to interactions retrieved from sentence-level processing. Black edges correspond to interactions retrieved from abstract level processing. The more sentences/abstracts describe an interaction between gene pairs the thicker the edge connecting them.

and the host gene *interleukin-4*. However, the co-occurrence methods incorrectly classified this pair as interacting, since these genes occur in the same sentence. This leads to drop in precision.

Support vector machines with edit and cosine kernel obtained a higher precision compared to the co-occurrence based approach. The precision obtained by the cosine kernel (71%) was significantly higher than the precision values of the co-occurrence and edit kernel approaches. Edit kernel, on the other hand, obtained more balanced precision and recall levels compared to the other methods.

Both edit kernel and cosine kernel operate on sentence-level. Therefore, they are not able to identify interactions whose descriptions cross sentence boundaries. The significantly higher number of true positive interactions retrieved by the abstract-level co-occurrence approach indicates the importance of the use of abstracts (or scopes wider than sentences) as context.

Figure 4 shows the literature mined and manually verified unique host-*Brucella* gene–gene interactions. A total of 46 unique interaction pairs are retrieved. 24 of these were identified using sentence-level processing. Abstract-level analysis enabled the retrieval of 22 additional unique interaction pairs

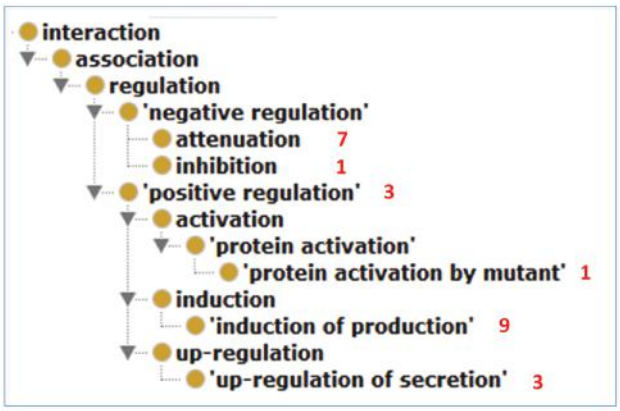


FIGURE 5 | The ontology hierarchy of literature mined INO interaction types. In total, six different INO interaction types were identified from this literature mining study. The number of interactions of a specific type is shown in red next to the interaction type. The 'induction of production' type is the most common type identified.

(**Figure 4A**). The identified host-*Brucella* gene–gene interactions are represented as a network, which consists of 20 *Brucella* genes and 25 host genes (**Figure 4B**). The interactions between host and *Brucella* gene pairs are represented as edges. The edges are weighed based on the number of sentences/abstracts that state the corresponding interaction. BLS and L7/L12 are the most connected *Brucella* genes, whereas IFNG and IRF1 are the most connected host genes.

Ontology Modeling of Host-*Brucella* Gene–gene Interactions

We used INO to analyze the types of interactions between the extracted host and *Brucella* genes. The results of this analysis are shown in **Figure 5**. In total, six different INO interaction types, all of which are sub-types of regulation, are identified from this literature mining study. The 'induction of production' type is the most common type identified. For instance, the sentence "The P39 and the bacterioferrin (BFR) antigens of *B. melitensis* 16M were previously identified as T dominant antigens able to induce both delayed-type hypersensitivity in sensitized guinea pigs and *in vitro* gamma interferon (IFN-gamma) production by peripheral blood mononuclear cells from infected cattle" (Al-Mariri et al., 2001) is an example sentence that describes an interaction of type 'induction of production' between pathogen and host genes. The sentence states that *Brucella* gene P39 is able to induce *in vitro* host IFN-gamma production.

While **Figure 4** provides concrete summary of the host-*Brucella* gene–gene interaction network, it is typical that each gene–gene interaction occurs under specific experimental condition(s). Without a specific condition, any host–pathogen interaction will not happen. Ontology provides an ideal platform to model and represent these gene–gene interactions under specific conditions. Below we provide two examples to illustrate how ontology-based gene–gene interactions work. These two examples include one retrieved from sentence level literature mining and another from abstract level literature mining. The ontology modeling uses the framework of the INO (Hur et al., 2015), the Ontology for Biomedical Investigations (OBI; Brinkman et al., 2010), and the Brucellosis Ontology (IDOBRU; Lin et al., 2011, 2015).

A host-*Brucella* gene–gene interaction based on literature mined sentence (Velikovsky et al., 2003) was modeled using ontology (**Figure 6A**). In this example, the mice were immunized with recombinant *Brucella* lumazine synthase (rBLS) administered with different adjuvants including incomplete Freund's adjuvant (IFA), monophosphoryl lipid A (MPA), and aluminum hydroxide gel (Al). The splenocytes were isolated from immunized mice and then re-stimulated with rBLS. Different cytokines (IFN-gamma, IL-2, IL-4, and IL-10) were produced by the splenocytes, indicating a mix of Th1 and Th2 response. This model represents the detail of the interactions between *Brucella* BLS and mouse IFN-gamma, IL-2, IL-4, and IL-10. This example is classified as another 'induction of production' interaction type (**Figure 5**), i.e., recombinant BLS induces the production of different proteins in splenocytes isolated from immunized mice.

Figure 6B provides another example of ontology modeling of the interaction between *Brucella* gene *wboA* and mouse protein Caspase-2, encoded by mouse gene Casp2, using the abstract content from the paper (Chen and He, 2009). *Brucella* mutant RA1, a mutant of wild type, virulent *B. abortus* strain 2308, lacks the *Brucella* gene *wboA*. RA1-infected RAW 264.7 mouse macrophage cell line cells had activated Caspase-2, which mediated apoptotic and necrotic cell death of RAW 264.7 cells (Chen and He, 2009). This example represents how *Brucella* gene *wboA* interacts with mouse Caspase-2. This example demonstrates the interaction type of 'protein activation by mutant' (**Figure 5**), i.e., a mutant of a gene infects mouse macrophages and activates the production of a mouse protein Caspase-2.

DISCUSSION

Using *Brucella* as an example pathogen, this study utilized literature mining and ontology analysis approaches to examine the interactions between host genes/proteins and *Brucella* genes/proteins. Since genes encode for proteins, our host-*Brucella* gene–gene interactions also include protein–protein interactions. Our approach identified 46 pairs of host-*Brucella* gene–gene interactions from the literature, and the ontology modeling analysis identified different types of interactions and provided deeper insights on how the host and *Brucella* genes/proteins interact at different experimental conditions.

One challenge in host–pathogen interaction literature mining is the difficulty in differentiating host genes and pathogen genes. In the current version of SciMiner and VO-SciMiner we did not use any of the name (longer description)-based identification results in the analysis. This is due to our manual evaluation of the preliminary results suggesting it is far more difficult to distinguish between host and pathogen genes using longer description protein names as they are more redundant than gene

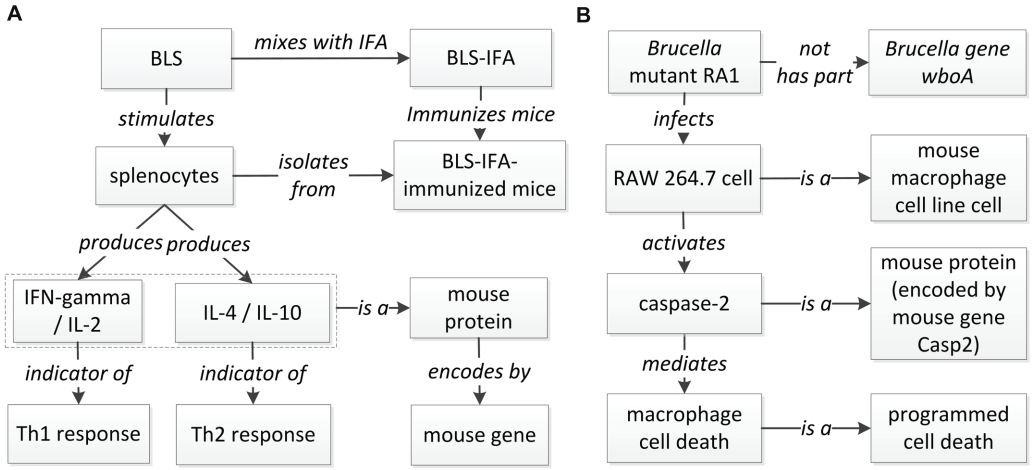


FIGURE 6 | Ontology modeling of literature-mined host-*Brucella* interaction types. (A) Ontology modeling of the gene interaction from the sentence "In addition, after *in vitro* stimulation with rBLS, spleen cells from BLS-IFA-, BLS-Al-, or BLS-MPA-immunized mice proliferated and produced interleukin-2 (IL-2), gamma interferon (IFN-gamma), IL-10, and IL-4, suggesting the induction of a mixed Th1-Th2 response" (Velikovsky et al., 2003). **(B)** Ontology modeling of the Casp2-wboA gene interaction using the abstract content from the paper (Chen and He, 2009).

symbols. For example, the protein name "Superoxide dismutase [Cu-Zn]" may represent a human/host gene name (SOD1 or SODC) or a *Brucella*/pathogen protein (SodC). In general, the gene names are more unique than the gene symbols; therefore, use of only short gene symbols resulted in decreased numbers of identified genes by the current versions of SciMiner and VO-SciMiner. We will examine these missed genes and further improve the sensitivity and accuracy of the gene name-based identification.

We investigated using co-occurrence and machine learning based methods for extracting host–pathogen gene–gene interactions. The co-occurrence based methods classify each pair of host and pathogen genes as interacting, if they occur in the same sentence/abstract. Therefore, they obtain high recall by retrieving all interacting pairs of genes. However, they also classify many gene pairs incorrectly as interacting, since not all co-occurring gene pairs are true interactions. This leads to drop in performance in terms of precision. The SVM classifiers with the dependency tree based edit and cosine kernels make use of the syntactic analysis of the sentences. These methods achieved higher precision compared to the co-occurrence based methods. To the best of our knowledge, there does not exist a large manually labeled host–pathogen gene–gene interaction data set. Therefore, the edit and cosine kernel based SVM classifiers were trained by using generic (intra-species) protein–protein interaction data sets. Training these classifiers with host–pathogen gene–gene interaction data might improve their performances. A drawback of most (if not all) currently available machine learning based interaction extraction methods is that they operate on sentence-level and therefore, are not able to identify interactions that cross sentence boundaries. As our sentence-level and abstract-level co-occurrence analysis revealed, many host-*Brucella* interactions span multiple sentences. These results suggest that developing text mining methods that operate on scopes wider than a sentence would be useful for extracting host–pathogen gene–gene interactions.

Our ontology modeling studies demonstrate its value in further identifying the nature and insights of host–pathogen gene–gene interactions. A simple gene–gene interaction may miss many details, especially in the setting of a host–pathogen interaction. A gene–(interaction type)-gene would provide more details since the interaction type could indicate how the two genes interact. The INO provides a way to classify hundreds of interaction keywords into logically defined interaction types under a hierarchical ontology setting (Hur et al., 2015). The usage of INO interaction types and its hierarchy allows us to detect the distribution of the interaction types from our literature mining study (**Figure 5**). INO-based modeling also provides a novel way to identify interaction types that are represented by multiple keywords in sentences (Özgür et al., 2015). Furthermore, ontology modeling of the mined sentences or abstracts provides a way to deeply identify the experimental setting where a host gene and a pathogen gene interact. Without such settings, detected host–pathogen interactions may not occur. Therefore, the ontology modeling is critical for our better detection and representation of the details of host–pathogen interaction mechanisms.

A promising future work is to use ontology modeling to identify possible types of patterns of how host and pathogen genes interact and apply such design patterns to guide our literature mining. For example, based on the ontology model of the 'protein activation by mutant' interaction type (**Figure 6B**), we may design a pattern-specific literature mining study. Specifically, a mutant represents a recombinant organism with the mutation of an internal gene. After a mutant is generated, a name is usually assigned to the mutant. As shown in **Figure 6B**, a pathogen mutant is often used in different experimental settings to infect a host and activate a host protein. Such a complex pattern is

difficult to retrieve using current literature mining strategies. For instance, a sentence often describes the relation between a mutant (instead of a pathogen gene) and a host gene. Based on the ontology-modeled pattern, we can first design a literature mining approach to identify all mutants and their corresponding pathogen genes; and based on the mutant-gene interaction, we can then infer the gene–gene interaction. Specific experimental conditions (e.g., host cell types) can also be mined using the ontology modeling. Literature mined and experimentally verified results can further be ontologically represented in an ontology such as the Brucellosis Ontology (IDOBRU; Lin et al., 2011, 2015).

Compared to model pathogens such as *Escherichia coli* and *Salmonella*, *Brucella* is a less studied pathogen. However, the results obtained from this study provide the first example of opportunities and challenges in the literature mining of the host–pathogen gene–gene interactions.

ACKNOWLEDGMENTS

This research was supported by grant R01AI081062 from the US NIH National Institute of Allergy and Infectious Diseases (to YH) and Marie Curie FP7-Reintegration-Grants within the 7th European Community Framework Programme (to AO). JH was partially supported by the University of North Dakota, Epigenomics Center of Biomedical Research Excellence (COBRE; NIGMS P20GM104360).

REFERENCES

Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., and Salakoski, T. (2008). All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 9:S2. doi: 10.1186/1471-2105-9-S11-S2

Al-Mariri, A., Tibor, A., Mertens, P., De Bolle, X., Michel, P., Godefroid, J., et al. (2001). Protection of BALB/c mice against *Brucella* abortus 544 challenge by vaccination with bacterioferritin or P39 recombinant proteins with CpG oligodeoxynucleotides as adjuvant. *Infect. Immun.* 69, 4816–4822. doi: 10.1128/IAI.69.8.4816-4822.2001

Arenas-Gamboa, A. M., Ficht, T. A., Kahl-Mcdonagh, M. M., and Rice-Ficht, A. C. (2008). Immunization with a single dose of a microencapsulated *Brucella melitensis* mutant enhances protection against wild-type challenge. *Infect. Immun.* 76, 2448–2455. doi: 10.1128/IAI.00767-07

Blaschke, C., and Valencia, A. (2002). The frame-based module of the SUISEKI information extraction system. *IEEE Intell. Syst.* 17, 14–20. doi: 10.1109/MIS.2002.999215

Brinkman, R. R., Courtot, M., Derom, D., Fostel, J. M., He, Y., Lord, P., et al. (2010). Modeling biomedical experimental processes with OBI. *J. Biomed. Semant.* 1(Suppl. 1), S7. doi: 10.1186/2041-1480-1-S1-S7

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., et al. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artif. Intell. Med.* 33, 139–155. doi: 10.1016/j.artmed.2004.07.016

Chen, F., and He, Y. (2009). Caspase-2 mediated apoptotic and necrotic murine macrophage cell death induced by rough *Brucella abortus*. *PLoS ONE* 4:e6830. doi: 10.1371/journal.pone.0006830

Corbel, M. J. (1997). Brucellosis: an overview. *Emerg. Infect. Dis.* 3, 213–221. doi: 10.3201/eid0302.970219

de Marneffe, M.-C., Maccartney, B., and Manning, C. D. (2006). "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC-06*, (Amsterdam: Elsevier).

Durmus, S., Cakir, T., Özgür, A., and Guthke, R. (2015). A review on computational systems biology of pathogen-host interactions. *Front. Microbiol.* 6:235. doi: 10.3389/fmicb.2015.00235

Erkan, G., Özgür, A., and Radev, D. R. (2007). "Semi-supervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, 228–237.

Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac. Symp. Biocomput.* 707–718.

Giuliano, C., Lavelli, A., and Romano, L. (2006). "Exploiting shallow linguistic information for relation extraction from biomedical literature," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, 401–408.

Halling, S. M., Peterson-Burch, B. D., Bricker, B. J., Zuerner, R. L., Qing, Z., Li, L. L., et al. (2005). Completion of the genome sequence of *Brucella abortus* and comparison to the highly similar genomes of *Brucella melitensis* and *Brucella suis*. *J. Bacteriol.* 187, 2715–2726. doi: 10.1128/JB.187.8.2715-2726.2005

Hsu, C.-N., Chang, Y.-M., Kuo, C.-J., Lin,-Y. S., Huang, H.-S., and Chung, I.-F. (2008). Integrating high dimensional bi-directional parsing models for gene mention tagging. *Bioinformatics* 24, i286–i294. doi: 10.1093/bioinformatics/btn183

Hur, J., Özgür, A., Xiang, Z., and He, Y. (2012). Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *J. Biomed. Semant.* 3:18. doi: 10.1186/2041-1480-3-18

Hur, J., Özgür, A., Xiang, Z., and He, Y. (2015). Development and application of an interaction network ontology for literature mining of vaccine-associated gene–gene interactions. *J. Biomed. Semant.* 6:2. doi: 10.1186/2041-1480-6-2

Hur, J., Schuyler, A. D., States, D. J., and Feldman, E. L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics* 25, 838–840. doi: 10.1093/bioinformatics/btp049

Hur, J., Xiang, Z., Feldman, E. L., and He, Y. (2011). Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network. *BMC Immunol.* 12:49. doi: 10.1186/1471-2172-12-49

Jelier, R., Jenster, G., Dorssers, L. C., Van Der Eijk, C. C., Van Mulligen, E. M., Mons, B., et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics* 21, 2049–2058. doi: 10.1093/bioinformatics/bti268

Joachims, T. (1999). "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, eds J. C. Christopher, B. S. Burges, and A. J. Smola (Cambridge, MA: MIT Press), 169–184.

Kumar, R., and Nanduri, B. (2010). HPIDB-a unified resource for host-pathogen interactions. *BMC Bioinformatics* 11:S16. doi: 10.1186/1471-2105-11-S6-S16

Lin, Y., Xiang, Z., and He, Y. (2011). Brucellosis ontology (IDOBRU) as an extension of the infectious disease ontology. *J. Biomed. Semant.* 2:9. doi: 10.1186/2041-1480-2-9

Lin, Y., Xiang, Z., and He, Y. (2015). Ontology-based representation and analysis of host-*Brucella* interactions. *J. Biomed. Semant.* 6:37. doi: 10.1186/s13326-015-0036-y

McDonald, R., and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics* 6(Suppl 1):S6. doi: 10.1186/1471-2105-6-S1-S6

O'Callaghan, D., and Whatmore, A. M. (2011). *Brucella* genomics as we enter the multi-genome era. *Brief. Funct. Genomics* 10, 334–341. doi: 10.1093/bfgp/elr026

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics* 17, 155–161. doi: 10.1093/bioinformatics/17.2.155

Özgür, A., Hur, J., and He, Y. (2015). "Extension of the Interaction Network Ontology for literature mining of gene–gene interaction networks from sentences with multiple interaction keywords," in *The 2015 International Workshop on Biomedical Data Mining, Modeling, and Semantic Integration*

(BDM2I 2015) workshop, eds S. Dezhao, F. Adam, T. Cui and S. Frank (Bethlehem: The International Semantic Web Conference) 12.

Özgür, A., Xiang, Z., Radev, D. R., and He, Y. (2011). Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J. Biomed. Semant.* 2(Suppl. 2), S8. doi: 10.1186/2041-1480-2-S2-S8

Rosinha, G. M., Myioshi, A., Azevedo, V., Splitter, G. A., and Oliveira, S. C. (2002). Molecular and immunological characterisation of recombinant Brucella abortus glyceraldehyde-3-phosphate-dehydrogenase, a T-and B-cell reactive protein that induces partial protection when co-administered with an interleukin-12-expressing plasmid in a DNA vaccine formulation. *J. Med. Microbiol.* 51, 661–671.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics* 6(Suppl. 1):S3. doi: 10.1186/1471-2105-6-S1-S3

Tekir, S. D. C., Cakir, T., Ardic, E., Sayilirbas, A. S., Konuk, G., Konuk, M., et al. (2013). PHISTO: pathogen-host interaction search tool. *Bioinformatics* 29, 1357–1358. doi: 10.1093/bioinformatics/btt137

Thieu, T., Joshi, S., Warren, S., and Korkin, D. (2012). Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* 28, 867–875. doi: 10.1093/bioinformatics/bts042

Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput. Biol.* 6:e1000837. doi: 10.1371/journal.pcbi.1000837

Tsai, R. T.-H., Sung, C.-L., Dai, H.-J., Hung, H.-C., Sung, T.-Y., and Hsu, W.-L. (2006). NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics* 7(Suppl 5):S11. doi: 10.1186/1471-2105-7-S5-S11

Velikovsky, C. A., Goldbaum, F. A., Cassataro, J., Estein, S., Bowden, R. A., Bruno, L., et al. (2003). *Brucella lumazine* synthase elicits a mixed Th1-Th2 immune response and reduces infection in mice challenged with *Brucella abortus* 544 independently of the adjuvant formulation used. *Infect. Immun.* 71, 5750–5755. doi: 10.1128/IAI.71.10.5750-5755.2003

Xiang, Z., Qin, T., Qin, Z., and He, Y. (2013). A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. *BMC Syst. Biol.* 7:S9. doi: 10.1186/1752-0509-7-S3-S9

Xiang, Z., Tian, Y., He, Y., and Others. (2007). PHIDIAS: a pathogen-host interaction data integration and analysis system. *Genome Biol.* 8:R150. doi: 10.1186/gb-2007-8-7-r150

Yin, L., Xu, G., Torii, M., Niu, Z., Maisog, J. M., Wu, C., et al. (2010). Document classification for mining host pathogen protein-protein interactions. *Artif. Intell. Med.* 49, 155–160. doi: 10.1016/j.artmed.2010.04.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

TECHNOLOGY REPORT
published: 06 October 2015
doi: 10.3389/fmicb.2015.01032

Cell scale host-pathogen modeling: another branch in the evolution of constraint-based methods

Neema Jamshidi[1,2]* *and Anu Raghunathan*[3]

[1] Institute of Engineering in Medicine, University of California, San Diego, La Jolla, CA, USA, [2] Department of Radiological Sciences, University of California, Los Angeles, Los Angeles, CA, USA, [3] Chemical Engineering Division, National Chemical Laboratory, Pune, India

Constraint-based models have become popular methods for systems biology as they enable the integration of complex, disparate datasets in a biologically cohesive framework that also supports the description of biological processes in terms of basic physicochemical constraints and relationships. The scope, scale, and application of genome scale models have grown from single cell bacteria to multi-cellular interaction modeling; host-pathogen modeling represents one of these examples at the current horizon of constraint-based methods. There are now a small number of examples of host-pathogen constraint-based models in the literature, however there has not yet been a definitive description of the methodology required for the functional integration of genome scale models in order to generate simulation capable host-pathogen models. Herein we outline a systematic procedure to produce functional host-pathogen models, highlighting steps which require debugging and iterative revisions in order to successfully build a functional model. The construction of such models will enable the exploration of host-pathogen interactions by leveraging the growing wealth of omic data in order to better understand mechanism of infection and identify novel therapeutic strategies.

Keywords: constraint-based model, host-pathogen, optimization methods, mathematical models, omics-technologies, tuberculosis, salmonella typhimurium, flux balance analysis

OPEN ACCESS

Edited by:
Tunahan Cakir,
Gebze Technical University, Turkey

Reviewed by:
Pinar Pir,
Babraham Institute, UK
Adil Mardinoglu,
Chalmers University of Technology, Sweden

***Correspondence:**
Neema Jamshidi,
Institute of Engineering in Medicine, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0412, USA; Department of Radiological Sciences, University of California, Los Angeles, BOX 951721, Los Angeles, CA 90095-1721, USA
neema@ucsd.edu

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *19 March 2015*
Accepted: *11 September 2015*
Published: *06 October 2015*

Citation:
Jamshidi N and Raghunathan A (2015) Cell scale host-pathogen modeling: another branch in the evolution of constraint-based methods. Front. Microbiol. 6:1032. doi: 10.3389/fmicb.2015.01032

Why Constraint-based Modeling for Host-pathogen Interactions?

Rudolph Virchow, a nineteenth century co-founder of pathology is credited with describing pathology as "physiology with obstacles" and specifying a "diseased state" as a quantitative deviation from normal function as a result of internal and external (i.e., environmental) influences (Virchow, 1958). Infections of a host by a pathogen can lead to acute and chronic pathological conditions. The process of infection by a pathogen can be viewed as a pathological process resulting from environmental stresses. These causal influences by the pathogen, onto the host, define the

Notations/Abbreviations: h, a host model; p, a pathogen model; hp, a host-pathogen model; BM,h, host biomass pseudo-reaction; BM,p, pathogen biomass pseudo-reaction; S, the stoichiometric matrix for a metabolic network; v, flux vector in a metabolic network; x, metabolite vector in a metabolic network; m, the number of unique, compartment specific metabolites in a stoichiometric matrix, i.e., |x|; n, the number of unique, compartment specific reaction fluxes in a metabolic network, i.e., |v|; R, rank of the stoichiometric matrix; N_r, size of the right null space; N_l, size of the right null space; α, biomass optimum of host model; β, biomass optimum of pathogen model; ϵ, simulation constant for setting lower bound minimum of biomass production.

capabilities of the host and its pathogen can be expressed as constraints on the metabolic capabilities of the host and pathogen (**Figure 1**).

The continued development of high-throughput technologies are enabling profiling of multi-cellular and multi-organism environments (Gawronski et al., 2009; Han et al., 2010; Pacchiarotta et al., 2012; McAdam et al., 2014). Such advances enable the detailed measurement of molecular changes occurring in host-pathogen interactions (Kim and Weiss, 2008; Stavrinides et al., 2008; Beste et al., 2013; Le Chevalier et al., 2014; Schoen et al., 2014; Chang et al., 2015; Henningham et al., 2015; Yao and Rock, 2015). Generation of these large datasets, in the context of the complexity of pathogenesis, highlight the need for systems based approaches for integration into a cohesive biologically interpretable framework (Durmus et al., 2015). Constraint-based modeling is an ideal approach for a systematic, integrated analysis of these data. The approach is based on well-defined stoichiometric biochemical transformations (including mass balance, reaction capacity, and directionality) and gene-protein-reaction (GPR) relationships allow mapping and integration of multiple, disparate data types. These methods can incorporate heterogeneous data-types that represent all hierarchies in the reductionist causal chain of an organism, thus enabling prediction of emergent properties (**Figure 1**). Additionally, constraint-based models circumvent the problem of over fitting data, which often plagues strictly statistical based methods. There exist a number of freely available tutorials and implementation tools and packages enabling the use of reconstructions for modeling, analysis, and simulation in the literature (Schellenberger et al., 2011; Liao et al., 2012; Ebrahim et al., 2013; Sadhukhan and Raghunathan, 2014; Palsson, 2015).

Where in the Tree Do Host-pathogen Models Lie?

Constraint-based modeling in metabolism has its roots in microbial organisms, but has progressively grown in the past decades to describe complex multi-cellular organisms and various processes (Reed and Palsson, 2003; Mo et al., 2007; Feist et al., 2009; Karlsson et al., 2011; Osterlund et al., 2012). There has been a continual, systematic growth and progression of constraint-based models which initially began as the formulation of a core biochemical network as a linear

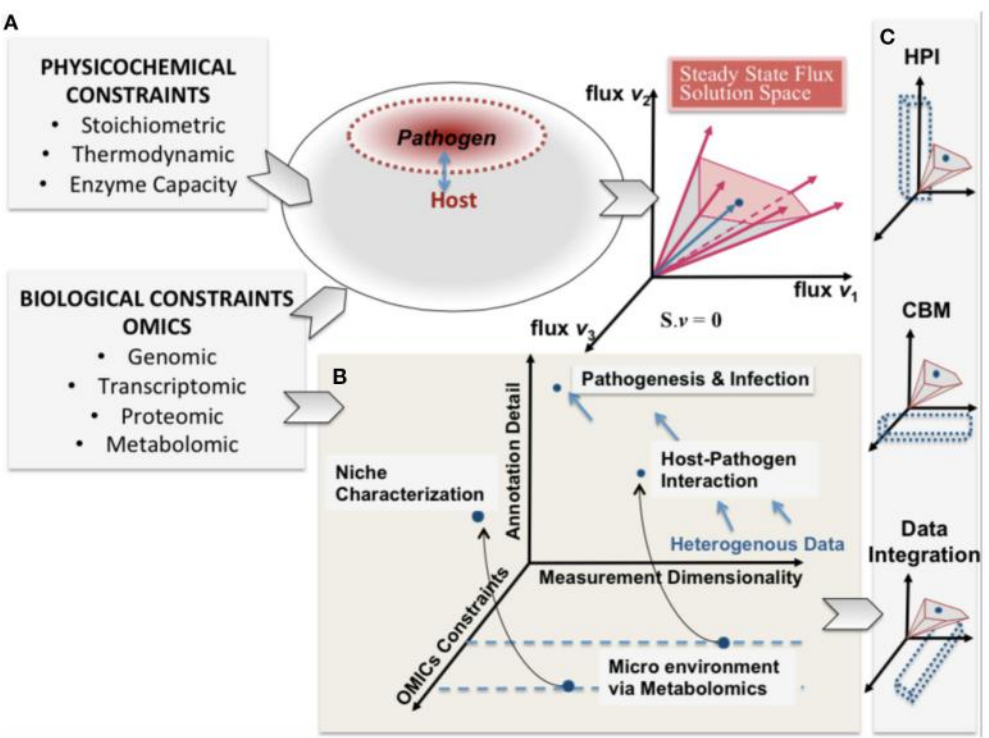


FIGURE 1 | A conceptual representation of integrating constraint-based modeling and omic data. The heterogeneity of omic data (biological constraints) and their integration is represented in parallel with the phenotypic solution space of the high dimensional host-pathogen model derived from physicochemical constraints. The degree of constraints represented will depend on the measurement capability and also define a reference set of behaviors that are feasible. **(A)** enumerates the heterogeneity of constraints for both host and pathogen and the resultant mathematically feasible and the potential biologically relevant solution space. In **(B)** pathogenesis and infection are shown from the perspective of 3 dimensions (i) omics constraints (also determined by experimental constraints) (ii) Annotation detail (based on existing legacy data) and (iii) the measurement dimensionality (also defining dimensionality of data). **(C)** shows that understanding host-pathogen interaction would be possible at multiple scales by integrating heterogeneous data/measurements and constraint-based modeling algorithms. The opportunity afforded by the legacies of high throughput omics experimentation and systems-level mathematical models would help understand the emergent host-pathogen interaction.

optimization problem (Papoutsakis, 1984; Fell and Small, 1986; Varma et al., 1993). Further incorporation of additional layers of biological information through GPRs, thermodynamic constraints, and various high throughput data have increased the scope of the models beyond small species metabolism, to multi-cellular, multi-compartmental organisms (Duarte et al., 2007; Mo et al., 2007; Herrgård et al., 2008; Lewis et al., 2010; Ahn et al., 2011; Bordbar et al., 2011; Chang et al., 2011; Saha et al., 2011; Mintz-Oron et al., 2012; Seaver et al., 2012; Wang et al., 2012; Pornputtapong et al., 2015). This evolution in the field has been accompanied by a growth in associated methodologies (Lewis et al., 2012) and new discoveries (Ellis et al., 2009; Ahn et al., 2011; Frezza et al., 2011; Thomas et al., 2014; Väremo et al., 2015). The importance of metabolism in understanding the process of infections and host pathogen relationships is increasingly being recognized (Han et al., 2010; Kafsack and Llinás, 2010; Pacchiarotta et al., 2012; Beste et al., 2013; Mcconville, 2014; Schoen et al., 2014; Yao and Rock, 2015). The cellular environment and repertoire of available metabolites is critical in characterizing and understanding how a pathogen interacts with and infects the host and constraint-based approaches can provide value insight into mechanisms of resistance and potentially new drug treatment targets (Chavali et al., 2008; Huthmacher et al., 2010; Bazzani et al., 2012; Kim et al., 2013; Shoaie and Nielsen, 2014; Tymoshenko et al., 2015).

In the "evolutionary tree" of constraint-based models, host-pathogen models lie between multi-cellular models, pathogen modeling, and new constraints/data integration approaches. There are now numerous exciting frontiers in the growth of these models, including the scope, incorporation of physicochemical constraints, multi-tissue, and multi-organism models (Cakir et al., 2006; Kümmel et al., 2006a,b; Beg et al., 2007; Duarte et al., 2007; Mo et al., 2007; Herrgård et al., 2008; Lewis et al., 2010; Ahn et al., 2011; Bordbar et al., 2011; Chang et al., 2011; Saha et al., 2011; Metris et al., 2012; Mintz-Oron et al., 2012; Seaver et al., 2012; Wang et al., 2012; Pornputtapong et al., 2015). Some of the challenges regarding model integration will be shared with related areas of multi-cellular constraint-based modeling, such as modeling microbial communities (Stolyar et al., 2007; Karlsson et al., 2011; Shoaie and Nielsen, 2014) and the development of new methods characterizing the interaction between cellular interactions between different species (Harcombe et al., 2014). Notable differences between host pathogen modeling and microbial community modeling include the specification of cellular objectives and constraints as well as differences in spatial compartmentalization (microbial community modeling will generally involve interaction through a shared extracellular space, whereas host pathogen models may interact through additional compartments; see below). We confine the scope of this work to focus on host-pathogen constraint-based modeling that entails the explicit integration of two genome-scale (or cell scale) constraint-based models. The purpose of this article is to describe a systematic methodology leading to successful integration of constraint-based host-pathogen models. Although there have been a relatively small number of actual host-pathogen (hp) models reconstructed to date, the existing studies have produced interesting results and have taken steps toward elucidating the pathway forward for future investigations (Raghunathan et al., 2009, 2010; Bordbar et al., 2010; Sadhukhan and Raghunathan, 2014).

The extracellular environment has an influential effect on the phenotype state and behavior of cells, thus pathogens have different biochemical phenotypes when inside the host versus outside the host and that the host cells will be affected in some manner by the pathogen and vice-versa. Many current experimental techniques enable characterization of these different states (Deatherage Kaiser et al., 2013). The generation of such data results in the technical challenge of simultaneous interpretation and analysis of genomic, proteomic, and/or metabolomics data of two independent, yet interacting organisms. The ability to derive meaningful interpretations of such data requires a computational setting which enables mapping and integrating data in a coherent format that further allows the data to be analyzed simultaneously, beyond simply looking at correlations or fitting presumed associations to an expected model. The constraint-based modeling framework affords a means to do so.

While there are a seemingly innumerable number of ways that pathogens have evolved to infect and reside their chosen host tissues and organs, in general terms there are few places these organisms can localize: intracellular, extracellular—interstitial, extracellular—intravascular, extracellular—transcellular, and "semi-open" spaces (e.g., the respiratory or alimentary tracts, etc.). In the constraint-based framework, there are three types of compartment based interactions between the host and pathogen (defined by the interaction boundary as defined by the pathogen's cell wall): extracellular, intracellular:cytosolic, intracellular:intra-organelle (**Figure 2**). Within the intracellular environment, there are multiple compartments that a pathogen may localize and life cycles of pathogens in some organisms reside in different compartments, depending on the stage of infection. These details are organism specific and are addressed on a case-by-case basis.

Reconstructing a Host-pathogen Constraint-based Model

Formulation of a Description of a Biochemical Network as a Constraint-based Optimization Problem

The formulation of metabolic network descriptions in terms of constraint-based modeling and relation optimization methods is rooted in applying the principle of mass conservation and thermodynamic constraints to these networks and has previously been described in detail (Fell and Small, 1986; Varma et al., 1993; Orth et al., 2010; Palsson, 2015). Integration of host-pathogen models requires two curated stoichiometric representations of metabolic networks, for which the minimum requirements are a stoichiometric matrix and a flux vector with upper and lower bounds,

$$S_h \cdot v_h = 0 \tag{1a}$$

$$v_h^{lb} \leq v_h \leq v_h^{ub} \tag{1b}$$

for the host and,

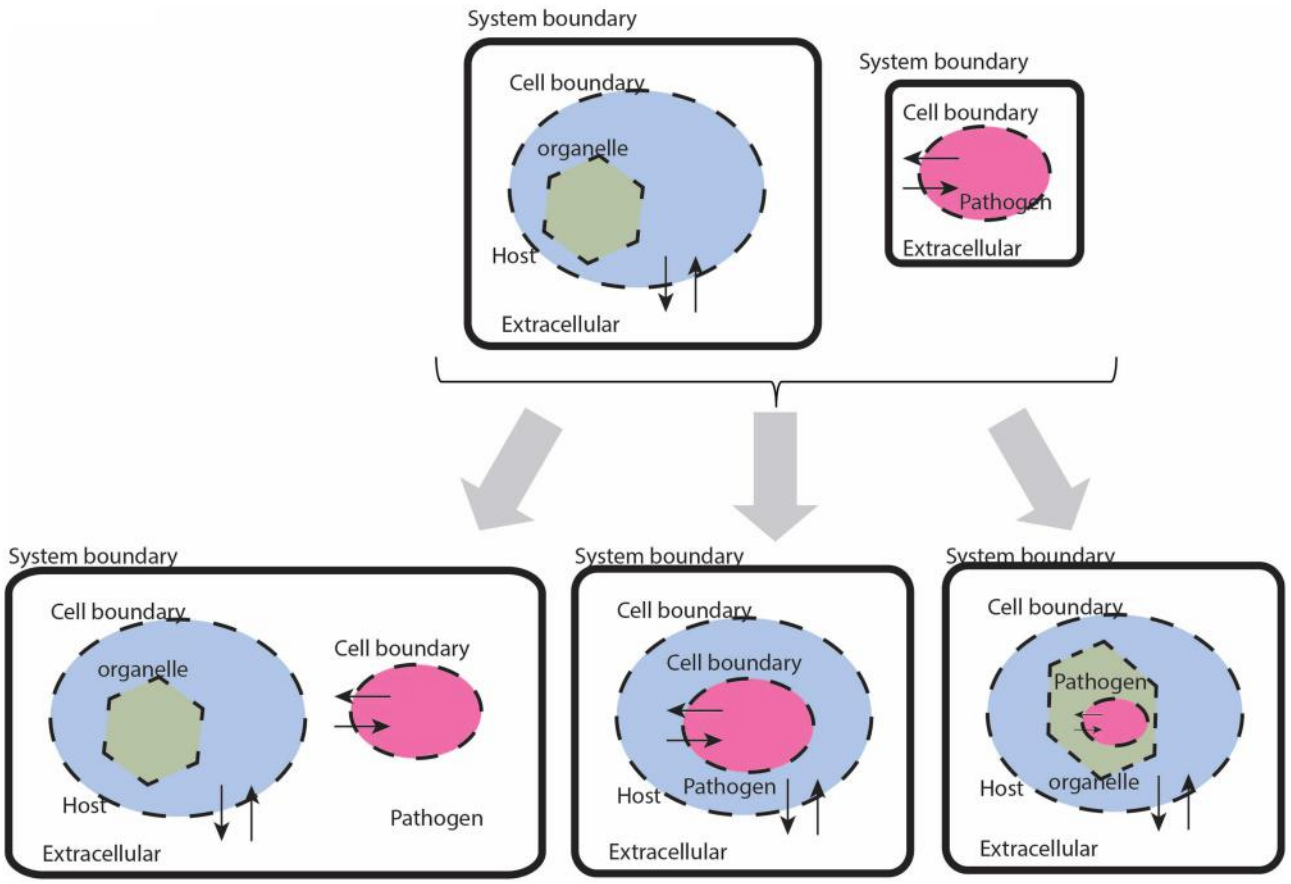


FIGURE 2 | Cartoon based schematic representation of different types of interactions between a host and pathogen model, with special attention to system boundaries. The system boundary is clearly delineated with a solid black line, whereas organism boundaries are dashed lines (the organelle boundary is represented with a thinner black line. Note that with this formulation, individual models will be required to have exchange reactions for every metabolite that has a trans-membrane transporter.

$$S_p \cdot v_p = 0 \tag{2a}$$

$$v_p^{lb} \leq v_p \leq v_p^{ub} \tag{2b}$$

for the pathogen, with $S_h \in R^{mh \times nh}$, $S_p \in R^{mp \times np}$, $v_h \in R^{nh}$, and $v_p \in R^{np}$ (see Notations/Abbreviations).

For host-pathogen modeling, Equations (1) and (2) are not applied under the strict steady state assumption, but rather along the lines of a quasi-homeostatic state for which we enforce mass conservation over a time scale of interest. With this consideration in mind, the calculation of interest is rarely a specific flux point, but rather a group of points reflecting a particular flux state (or a region within the right null space) corresponding to a particular phenotype that can be differentiated from other qualitatively different flux states. Identification of such regions often may not require the specification of a metabolic objective function, in which case non-objective based methods, such as sampling, may be appropriate (Savinell and Palsson, 1992; Barrett et al., 2006; Schellenberger and Palsson, 2009; Bordel et al., 2010).

Pre-existing curated, functional models are a necessary but not sufficient requirement for building an hp model. Even if two models are well posed, integration of the two may result in discrepancies as a result of multiple factors including,

- Error ranges in experimentally derived values (such as biomass components).
- Incorporation of data from different experimental conditions that may not be consistent with one another from a mass balance or thermodynamic perspective.
- Limitations in biological scope of each respective model.
- Lack of knowledge about the true or underlying biological objectives.

Additional, important considerations to be made when transitioning from the analysis of an isolated pathogen to a host-pathogen model include, simulating different conditions with different data sets, simulating the same species under different states versus different species under similar conditions, and specification of the conditions in which gene lethality knockout/knockdown studies or drug sensitivity screens are performed and their applicability to host-pathogen infectious states. These issues highlight the need for a systematic methodology for integrating host-pathogen models.

Constraint-based host-pathogen modeling can be viewed as a generalizable, systematic, multi-tiered process with iterative sub-steps (**Figure 3**). Each step includes multiple sub-steps that require simulations or calculations to be performed, often in an iterative fashion. A systematic approach for building and testing the models during the integration process will help make the debugging process more transparent and the more directed identification of potential problems.

Step 1. Pre-integration Model Check

This initial step serves as a "sanity check" to avoid problems during the subsequent integration components of the study. Although current standards for building curated network reconstructions generally require critical quality control/quality assurance steps to avoid spurious behavior from ill-posed models, prior to integration, there are a number of tests that must be completed for each model to confirm the models have been constructed and formulated appropriately.

1.i Check mass balances ("No free lunch"). Well curated models should be free of errors that may lead to violation of mass conservations constraints. However prior to integration, each model should be tested to confirm this, i.e., all uptake exchange reactions should be closed and flux variability analysis (FVA) (Mahadevan and Schilling, 2003) should be performed on the entire model, in order to confirm that there is no *net* production of *any* metabolite, when no substrates are available for uptake. In the toy model depicted in **Figure 4**, it is clear that if the substrates for the host and pathogen are not available (F_e, A_e, and D_e), then none of the secreted compounds (B_e, X_e, E_e, Q_e) can be produced.

1.ii Identify boundary points. The simplest approach for identification of the boundary points for a model is through FVA. Although this step can technically be included in the Functionality Test Suite, FVA is such a useful tool for debugging and initial assessment of models, that it is judicious to include this as a mandatory step in the model integration protocol. Under general uptake conditions (that are still biologically and thermodynamically feasible), FVA is performed with subsequent calculation of the flux spans. This assessment will enable the determination of the ranges of all reactions and the potential identification of "closed" reactions, any unbounded reactions, etc.

1.iii Functionality test suite. Prior to integration there should be a pre-defined set of simulation condition(s) and reaction optimizations in order to test and confirm desired functionality of the model (Duarte et al., 2007); this set of reactions comprise the Functionality Test Suite (FTS). The FTS can contain any number of desired tests and simulations to ensure appropriate physiologic behavior of the model, examples include biomass production under different growth conditions, specific gene knockout lethality experiments, inability to growth under specified conditions, or any other appropriate test that would evaluate the physiological/biological characterization of the model or the underlying mathematical definition.

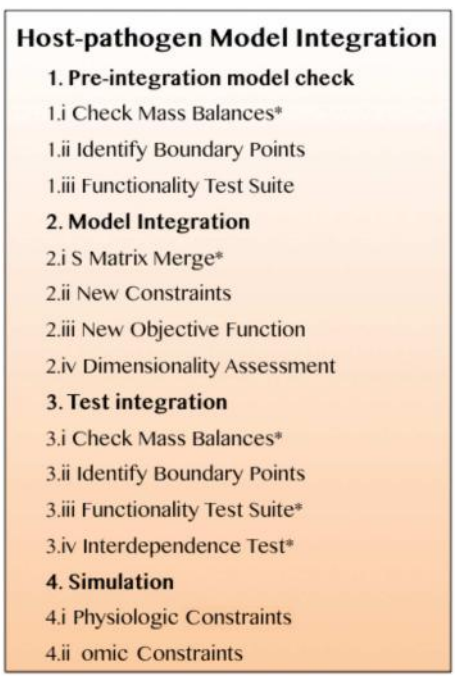


FIGURE 3 | A systematic procedure for successful, functional integration of a constraint-based host-pathogen model. Details are described in the main text. The asterisks identify steps that require iterative revisions if the models fail the corresponding test (see *Iteration/revision checkpoints in the main text).

Step 2. Model Integration

Although stoichiometric matrix integration of two models is trivial from a technical standpoint, the functional integration of a simulation-capable host and pathogen network reconstruction is a non-trivial process. The panels in **Figure 4** provide a concrete illustration of the integration of two toy models.

2.i S matrix merge. The stoichiometric matrices are joined through a compartment specific, row wise-merge (**Figure 4**). Generally compartment specific reactions (i.e., the compartment in which nutrients are directly exchanged between the host and pathogen) will not be shared between the host and pathogen model, however it is important to confirm this when constructing the new stoichiometric matrix.

$$m_{hp} < m_h + m_p \quad (3)$$

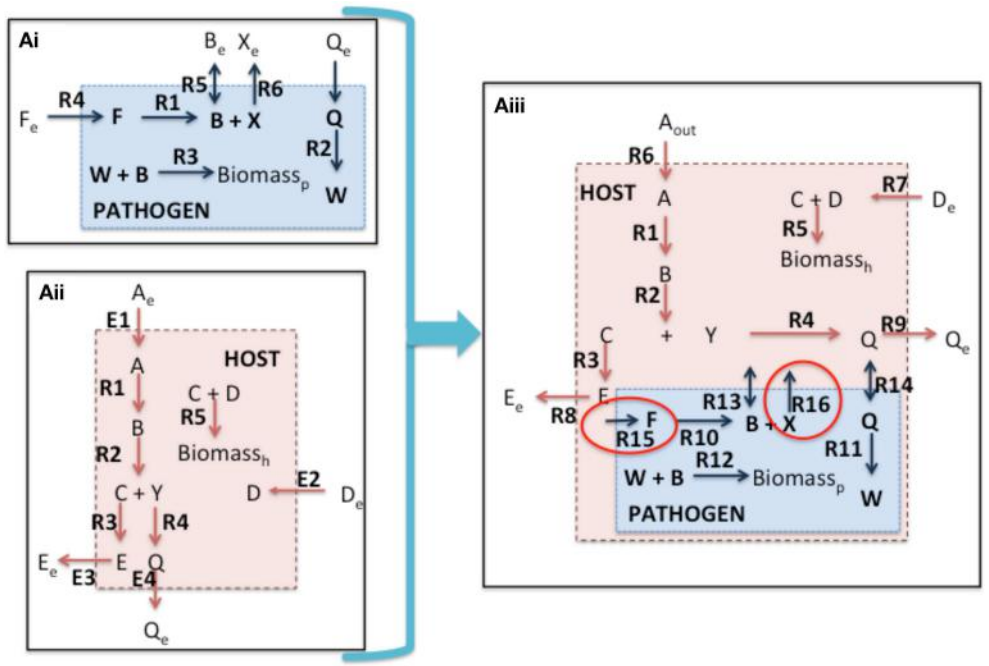
$$n_{hp} \approx n_h + n_p \quad (4)$$

Note that Equation 3 is defined by an inequality, whereas Equation (4) is an approximation. The degree of integration and subsequent complexity of the interactions between the models is dependent on the number of metabolites that overlap between the two organisms. If the organisms do not share any metabolites ($m_{hp} = m_h + m_p$), then integration of the two models will not result in any novel predictions. On the other hand, the number of reactions in the combined network may be equal to, less than, or greater than the sum of

the two individual models. In toy model integration depicted in **Figure 4**, $m_p = 9$, $m_h = 11$, and $m_{hp} = 18$, satisfying the Equation (3) inequality. For the toy model, Equation (4) is an equality, since the number of reactions in the combined model is equal to the sum of the individual models.

2.ii New constraints. Integration of two models includes the introduction of additional constraints that will make the simulation environment context specific and more representative of the actual biological environment.

- Nutrient availability and demand. These constraints are the most simple to implement and should provide strong coupling between the host and pathogen. In addition to biomass (growth and non-growth associated constraints), additional condition dependent constraints can be incorporated, for example demands on micronutrients, sequestration of metabolites, etc. (Rodriguez et al., 2002; Pan et al., 2010; Weiss and Schaible, 2015). For example in the toy model (**Figure 4**), further curation may be



B

P	INTRACELLULAR			TRANSPORT				EXCHANGE			
	R1	R2	R3	R4	R5	R6	R7	E1	E2	E3	E4
F	-1	0	0	1	0	0	0	0	0	0	0
F_e	0	0	0	-1	0	0	0	-1	0	0	0
B	1	0	-1	0	1	0	0	0	0	0	0
B_e	0	0	0	0	-1	0	0	0	-1	0	0
X	1	0	0	0	0	1	0	0	0	0	0
X_e	0	0	0	0	0	-1	0	0	0	-1	0
Q	0	-1	0	0	0	0	1	0	0	0	0
Q_e	0	0	0	0	0	0	-1	0	0	0	-1
W	0	1	-1	0	0	0	0	0	0	0	0
LB	0	0	0	-∞	-∞	-∞	-∞	-5	-5	-5	-5
UB	∞	∞	+∞	∞	+∞	+∞	+∞	0	0	0	0

FIGURE 4 | Continued

C

H	INTRACELLULAR					TRANSPORT				EXCHANGE			
	R1	R2	R3	R4	R5	R6	R7	R8	R9	E1	E2	E3	E4
A	-1	0	0	0	0	1	0	0	0	0	0	0	0
A_e	0	0	0	0	0	-1	0	0	0	-1	0	0	0
B	1	-1	0	0	0	0	0	0	0	0	0	0	0
C	0	1	-1	0	-1	0	0	0	0	0	0	0	0
D	0	0	0	0	-1	0	1	0	0	0	0	0	0
D_e	0	0	0	0	0	0	-1	0	0	0	-1	0	0
E	0	0	1	0	0	0	0	1	0	0	0	0	0
E_e	0	0	0	0	0	0	0	-1	0	0	0	-1	0
Y	0	1	0	-1	0	0	0	0	0	0	0	0	0
Q	0	0	0	1	0	0	0	0	1	0	0	0	0
Q_e	0	0	0	0	0	0	0	0	-1	0	0	0	-1
LB	0	0	0	0	-∞	-∞	-∞	-∞	-∞	-10	-10	-10	-10
UB	∞	∞	∞	∞	∞	∞	∞	∞	∞	0	0	0	0

D

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	E1	E2	E3	E4
A	-1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A_e	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	-1	0	0	0
B	1	-1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
C	0	1	-1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
D_e	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	-1	0	0
E	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
E_e	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	-1	0
Y	0	1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
Q_e	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0	-1
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0
X_e	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F_p	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0
B_p	0	0	0	0	0	0	0	0	0	1	0	-1	-1	0	0	0	0	0	0	0
X_p	0	0	0	0	0	0	0	0	0	1	0	0	0	0	-1	0	0	0	0	0
Q_p	0	0	0	0	0	0	0	0	0	0	-1	0	0	-1	0	0	0	0	0	0
W_p	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0
LB	0	0	0	0	-∞	-∞	-∞	0	0	-∞	-∞	-∞	-∞	0	0	-∞	-∞	-10	-10	-10	-10
UB	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	0	0	0	0

FIGURE 4 | Continued

FIGURE 4 | Integration of toy a host cell model with an intracellular pathogen model. (A) depicts a cartoon schematic of a pathogen model, host model, and integrated host-pathogen model with the corresponding stoichiometric matrices for each of the models (**B** corresponds to **Ai**, **C** corresponds to **Aii**, and **D** depicts the stoichiometric matrix for the hp network in **Aiii**). Note that when the pathogen "infects" the host the transporters for metabolites B and Q enable usurpation of host resources and will consequently limit the biomass construction capabilities of the host (potentially the pathogen as well, depending on the size of the demand). In the provided example, metabolites F and X are not within the intracellular environment of the host, thus R10, R15, and R16 will not be able to carry a flux. In spite of this however, since there is a transporter for metabolite B, the pathogen biomass can still be produced even though R10 will not be able to carry a flux. It is also possible that metabolite F and/or X actually *are* available in the host, but that the particular metabolites were outside the scope of the reconstruction at that time. In this case, the host model can be updated to include the relevant reactions that would make the metabolites available within the intracellular environment. The multiple points within the protocol that would allow for evaluation of the appropriateness of including additional reactions during the iterative revisions, particularly Steps 3.iii, 3.iv, and 4.i. Intracellular organelles are not described in this toy example, however if the pathogen infects the host and resides within a particular organelle within the host cell, the procedure would be the same. Note that the exchange reactions are not explicitly illustrated within the figures, but the columns are present in the stoichiometric matrices.

needed in order to identify the appropriate bounds for the intracellular pathogen uptake conditions as well as any potential new demands on available host nutrients (not depicted in this example).

- Coupling constraints. The host and pathogen networks will interact by virtue of the compartment specific shared metabolites. However, physiologically, the infection of a host by the pathogen frequently results in additional interdependencies between the two species, such as competition for a shared resource. Coupling constraints are the mathematical relationships formalizing the explicitly link between the host and pathogen models together as a constraint. This relationship may take the form as an interaction between two molecules, concordant activity between two enzymes, or some other biological process. For example,

$$v_i^h + / - v_j^p = \alpha_k \quad (5)$$

in which α_k is a physiologic constant or data dependent variable (e.g., protein production rates, mRNA expression, etc.). Non-unity coefficients can be added to the reactions, if there is known to be a fixed, stoichiometric balance between the two (or more) reactions. Depending on the type of relationship represented, this relationship can be expressed as a continuous flux based problem, or a discontinuous/discrete problem; the latter would require formulation as a Mixed Integer Linear Programming (MILP) problem (Burgard et al., 2001; Phalakornkule et al., 2001; Pharkya et al., 2004; Kumar and Maranas, 2009). In the case of hp models, MILP constraints may be used to express conditionally active reaction constraints. For example in the toy model depicted in **Figure 4**, if pathogen growth (i.e., biomass production, **Figure 4Aiii**, R12) were to only occur if the host cell would take up a particular metabolite (e.g., metabolite D, **Figure 4Aiii**, R7).
- State changes. To date methodologies for representing changes in infectious states during an infectious cycle or a pathogens life cycle have been represented as discrete, independent simulations. Depending on the type of data that is available, context specific models can be constructed for each different state or alternatively, conditional, state dependent constraints MILP constraints can be defined.

2.iii New objectives. Flux balance analysis is an optimization problem and while there are formulations of the constraint-based modeling problem that do not require the definition of a metabolic objective to be optimized (Lewis et al., 2012), the incorporation of an objective function to be maximized or minimized is often of great utility, since it enables more specific predictions to be made by reducing the size of the steady state solution space (right null space). The definition and identification of objective functions is an area of great importance in these models (particularly mammalian cell models) that is a very rich area for exploration and in need of further development in the current literature (Khannapho et al., 2008; Schuetz et al., 2012; Shoval et al., 2012; Szekely et al., 2013). The flexibility in designing cellular objectives to tailor hp specific responses is critical for achieving success with this approach. The biomass objective function has been discussed in great detail and is generally considered in terms of two general components: a growth associated component (accounting for biomass constituent components) and a non-growth associated component (Feist and Palsson, 2010). The biomass reaction can be treated as a constraint on the system or as a prediction to be made by the model as a means to validate a network reconstruction (Price et al., 2004). Since the growth of the solution space can increase dramatically when two models are merged, defining lower bound constraints on growth associated and non-growth associated biomass functions for the host or pathogen is a practical necessity in order to calculate meaningful results. Organism specific objectives may be developed from the new constraints that are defined or identified experimentally.

The specification of appropriate objective functions requires detailed understanding of pathogen physiology and host pathogen interactions. These can be separated into two general categories, single objective and multi-objective problems (**Figure 5**). Examples of potential objective functions include but are not restricted to, the (pathogen) biomass pseudo-reaction, iron acquisition (Ratledge and Dover, 2000; Nairz et al., 2015), lactate dehydrogenase levels as a indicator level of cytotoxicity (Korzeniewski and Callewaert, 1983; Decker and Lohmann-Matthes, 1988), enterotoxin production, pathogen specific metabolite production (Glickman et al., 2000; Takayama et al., 2005), reactive oxygen species minimization

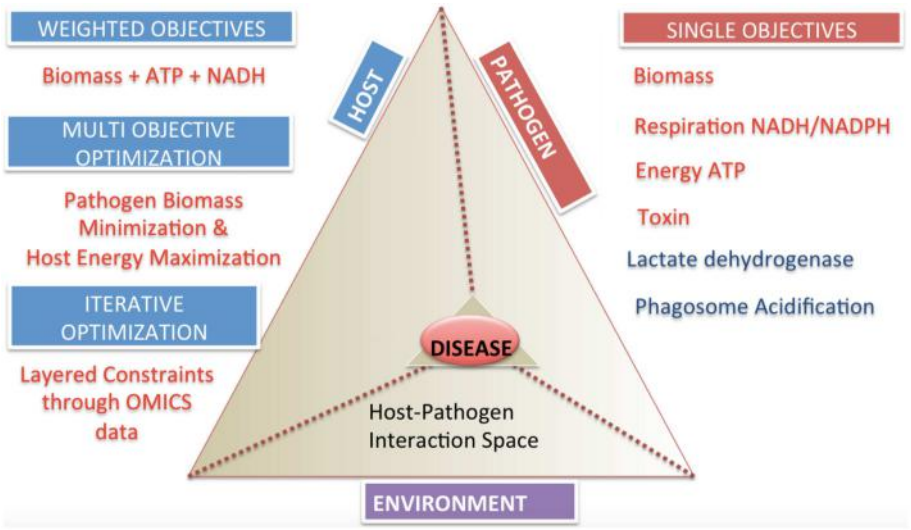


FIGURE 5 | Categories and classifications of objective functions in host-pathogen models. The host-pathogen interaction pyramid is shown that integrates host, pathogen and environment to result in the diseased state phenotype. The diseased state can be queried with the correct formulation of objective functions as discussed for the three components delineated here. The two sides of the triangle represent the host and the pathogen and the connecting side represents the environment or niche. The sides converge on the vertex of the prism reflecting the lethal disease state. The space outside the host pathogen interaction prism lists objectives and their classification. Single objectives help define pathogen or host state, while multi-objectives or weighted objective functions allow definition of complex phenotypes.

(Brynildsen et al., 2013), and other critical minerals and metabolites.

Multi-objective functions are more complex, but may reflect a more accurate representation of the biology (Gianchandani et al., 2008; Schuetz et al., 2012; Zakrzewski et al., 2012). The practical challenge is knowledge of the adequate data to specify these objectives.

- Weighted objectives. New objective functions can be constructed from the linear combination of reactions representing cellular demands and requirements. By combining different reactions together to generate "compound" or weighted objectives, more complex behavior can be captured. The obvious weakness of this approach is that the stoichiometric coefficients are fixed for the different components, thus this approach is only applicable in situations in which fluxes (or metabolite production/consumption) occur in fixed ratios with one another (as in biomass).
- Bi-level optimizations across host-pathogen boundaries. Bi-level optimization algorithms designed for bioengineering and evolutionary objectives (Burgard et al., 2003; Zomorrodi and Maranas, 2012) can be extended and applied to understand the dynamics across host and pathogen during an interaction. Depending on the experimental conditions, this may include optimization of pathogen biomass within the host. For example there may be competing objective functions, as in the case of maximization of pathogen biomass and host biomass concurrently or in diametric opposition, i.e., maximization of pathogen biomass with minimization of host substrate availability (either through minimization of pathogen transport uptake or host transport uptake).
- Multi-level optimization. Although, computationally intensive, multi-objective optimization (Zakrzewski et al., 2012; Zomorrodi et al., 2014) can enable a more accurate representation and in turn more accurate mathematical simulation of the host-pathogen interaction.
- Step wise algorithmic multi-objectives i.e., sequential optimizations that apply additional constraints at each iteration. Iterative optimizations are approach for including multilayered omic or physiological constraints allow to be added in order to asses hp behavior in varying environments or host niche's (D'Huys et al., 2012). Such approaches also support the integration of heterogeneous data types. A limitation of this approach is that the optimization is order dependent, and thus may be a more valuable tool for assessing the effects of different constraints as opposed to a more physiological objective.

2.iv Dimensionality assessment. Dimensionality assessment of the network includes determining the size of the network, including the number of metabolites and reactions, as well as the size of the "functional" space of the network, such as the right and left null spaces. These components can be directly calculated from m, n, and the rank of the new stoichiometric matrix S_{hp}. These quantities can be used to calculate the size of the right and left null spaces ($N_r = n–R$ and $N_l = m–R$). These simple calculations allow assessment of the dimensionality of the new model (in terms of number of components and reactions, as well as the steady state solution space), which will assist in debugging and interpretation of

TABLE 1 | Descriptive summary of the toy models.

	Pathogen	Host	Host-pathogen
Number of metabolites	9	11	18
Number of reactions	11	13	20
Right null space dimension	2	2	3
Left null space dimension	0	0	1
Rank	9	11	17
Mean betweenness centrality	0.11	2.64	4.33

The sizes of the stoichiometric matrices and the respective right and left null spaces. The steady state flux states reside in the right null space (calculated from applying mass conservation constraints). The left null space size describes the number of "conserved" metabolic moieties. In the case of the toy model, the left null space compound is the metabolite that cannot be imported into the pathogen, because the host model does not import or metabolize it.

calculated results and simulations (notably Steps 3 and 4). **Table 1** summarizes these results for the toy models described in **Figure 4**. Knowledge of the right null space in particular is useful when debugging potential problems and interpreting simulation results. Integration of the two models results in an increase in the steady state solution space (i.e., at least 1 new independent metabolic pathway) as a result of the integration from the host and pathogen. The left null space contains the conserved chemical moieties within a network (Famili and Palsson, 2003; Sauro and Ingalls, 2004). The size and contents of the left null space can be used to understand how metabolites may pool together based on network structure and often provides functional insights (Famili and Palsson, 2003; Thomas et al., 2014).

Additional graph theoretic measures can be calculated (Girvan and Newman, 2002; Estrada and Rodríguez-Velázquez, 2005; Fatumo et al., 2011), although their utility in assessment of functional characteristics and trouble-shooting in the context of hp model construction is currently limited.

Step 3. Integrated Host-pathogen Testing

On the surface, integration of two models is a trivial step given the general simplicity of the basic formulation of constraint-based models. The initial technical challenge is to identify the overlapping set of metabolites and corresponding abbreviation mappings between the host and pathogen metabolites.

Although there are laudable efforts to use standardized nomenclature (Radrich et al., 2010; Dräger and Palsson, 2014), a persistent challenge in the field is the use of different abbreviations and nomenclature, which has often required dedicated efforts to reconcile multiple versions of network reconstructions (Herrgård et al., 2008; Thiele and Palsson, 2010). Fortunately, however, for host pathogen models, every metabolite within the two models does not need to be compared, but rather just the boundary metabolites, which are generally a fraction of the total number of metabolites in a model. This is relatively straightforward through the comparison of abbreviations, if the reconstruction has been appropriately annotated [e.g., molecular formula, SMILES (Weininger, 1988), ChEBI (Degtyarenko et al., 2008), etc.]. Once the shared metabolite complement is identified, the stoichiometric matrices can be merged (**Figure 4**). However, "blind" integration without proper quality control/quality assurance and test conditions in place, the results will be difficult and quickly overwhelming to interpret.

The first three sub-steps for Step 3 are similar to Step 1. Depending on the type and complexity of new constraints that are applied to the integrated host-pathogen model, there are situations that may introduce behavior that violates mass conservation, thus it is necessary to confirm that no "free metabolites" are produced. For situations in which the pathogen is an intracellular organism, the test needs to be applied to the host-pathogen model, as well as the isolated pathogen, within the host.

3.i Check mass balances. See Step 2. Model Integration, 2.i and **Figure 3**, 2.i.

3.ii Identify boundary points. Identification of the right null space boundary points through FVA of the host-pathogen draft model will permit a detailed, yet global view of the capabilities of the combined host-pathogen and enable comparisons to the individual organisms (Step 1.ii). This comparison may identify reactions or constraints that may require revisions to be made. For example, upper bounds constraints may need to be increased if the combined model enables the pathogen to exceed the upper limit of some reactions in comparison to the isolated organism. In the case of the toy model illustrated in **Figure 4** (integrated host pathogen model), if host's intracellular environment is much richer than the "open" environment for the pathogen and in the infected state, R4 >> R12 (**Figure 4Aiii**), then the upper bound of R12 may need be increased in order to permit a larger potential rate of biomass accumulation.

3.iii The functionality test suite. The functionality test suite of the combined host-pathogen model will also enable a basis for comparison with 1.iii and assist subsequent analyses (Step 4). Note that the FTS for the individual host and pathogen models may not be identical to the hp set of test reactions, since the metabolic network capabilities of the host and pathogen will not be identical in the infected versus uninfected states.

3.iv Interdependence test This test requires identifying objective functions that are expected to influence or be influenced by the coupling between the host and pathogen. The biomass function is a very good candidate for such tests, as it is connected to many different pathways within each respective organism, and subsequently more likely to be directly connected to the host (or pathogen). The biomass pseudo-reaction, however, is not the only possible objective to test and other cellular/metabolic functions may be of utility, such as ATP production, oxidative phosphorylation, or constraints on secretion/uptake of particular metabolites (Schuetz et al., 2007, 2012; Khannapho et al., 2008; García Saánchez and Torres Sáez, 2014).

The interdependence test involves two steps,

a. Calculate the optimal host biomass production in the host-pathogen model, then fix the lower bound of the

host biomass reaction to a specified value (1-ε_1) and then optimize for the biomass of the pathogen:

$$\text{For } \alpha_1 = \max(v_{hp}^{BM,h}),$$
$$\text{set : lower bound}(v_{hp}^{BM,h}) \geq (1-\varepsilon_1)\alpha_1$$
$$\max (v_{hp}^{BM,p}) = \beta_2$$

b. Calculate the optimal pathogen biomass production in the host-pathogen model, then fix the lower bound of the pathogen biomass reaction to a specified value (1-ε_2) and then optimize for the biomass of the host:

$$\text{For } \beta_1 = \max(v_{hp}^{BM,p}),$$
$$\text{set : lower bound}(v_{hp}^{BM,h}) \geq (1-\varepsilon_2)\beta_1$$
$$\max (v_{hp}^{BM,h}) = \alpha_2$$

Comparison of α_1 to α_2 as well as β_1 to β_2 provides an indication of the degree of coupling between the two models. If $\alpha_1 \approx \alpha_2$ and $\beta_1 \approx \beta_2$, then there is no significant coupling between the two models. Conversely, if these values are significantly different from one another then there is evidence of interaction between the models on a metabolic level. It is more common to have uni-directional coupling between the models, often in favor of the pathogen, i.e., $\beta_1 \approx \beta_2$ and $\alpha_1 > \alpha_2$ due to usurpation of host resources by the pathogen. The ε coefficients are empiric, simulation based parameter whose value will vary depending on the specific organism, the biomass composition, and the media growth conditions. The "ideal" ε will be large enough to force the consumption of metabolites and resources required to produce biomass, but small enough not to introduce a significant bias in the flux state. When the coefficient ε is equal to 0, then the interdependence test is equivalent to a stepwise optimization comparison. Generally the coefficient ε is small, typically 0.01–0.1, when the biomass function is used. A phase portrait analysis (Edwards et al., 2002) may be useful in assessing and determining an appropriate ε value. Since ε is a specified value, the degree of coupling between the host and pathogen can be titrated to a certain degree. Note that since the growth rates of the host and pathogen may be very different from one another, then ε_1 and ε_2 may be different from one another.

Since the corners of the right null space generally become increasingly acute as the size of the model increases, when the biomass is fixed at the optimum level there is a dramatic decrease in the available alternative solutions. However when this constraint is relaxed even by a small amount, the number of alternative solution points dramatically expands; thus in order to assess robust coupling between the host and pathogen, generally a non-zero ε should be chosen.

For example in the toy model depicted in **Figure 4**, the pathogen biomass function is dependent on substrates provided by the host. If the uptake of metabolite A (**Figure 4Aiii**, R6) is unbounded (or not known to have any constraint), then the intra-cellular reproduction of the pathogen is not significantly constrained and independent of the host. However, if the host's uptake of metabolite A is limited, then the pathogen's growth rate will be limited. A common source of error and potential difficulty during the integration of a host and pathogen model is for the pathogen biomass production rate lower bound to be set above the availability of the particular metabolite (i.e., either the host uptake constraints or the host to pathogen transport reactions), which results in a non-functional model. In these cases, the data used for defining the constraints must be re-evaluated and either the constraints would need to be revised or there additional reactions would need to be added to provide alternative routes for availability of the requisite metabolite(s).

Step 4. Simulation

The type of simulation of interest is principally dependent on (1) the type of data available, (2) the biological organism of interest, and (3) the data available to validate or test the simulations. Due to the broad scope and scale of the realm of possible simulations, it is not practical to specify a list of calculations that can be applied for every condition. The purpose of this step is to assist in bridging the construction of the model to a meaningful use of the model in the subsequent analysis steps. A common characteristic of the simulation stage however involves evaluation steps and the question of how to reconcile inconsistent results between the model simulations and experimental observations. Suffice it to say that the use of integrated omic data is one of the most successful aspects of constraint-based modeling and there are a number of growing methods being developed for incorporating genomic sequence, transcriptomic, proteomic, and metabolomic data; interested readers are referred to available review articles outlining some of these methods (Blazier and Papin, 2012; Lewis et al., 2012; Wang et al., 2012; Machado and Herrgård, 2014; Robaina Estévez and Nikoloski, 2014). For the purpose of organization and simplifying the debugging process, the simulation tests can be classified into two general areas,

4.i Physiological constraints. Simulations validating (or invalidating) predictions of the model using available physiologic data sets.
4.ii Omic constraints. Simulations validating (or invalidating) predictions of the hp model through omic data sets.

*Iteration/revision checkpoints

"Failure" of specific steps in the protocol (**Figure 3**) requires an iterative adjustment to be made through revision of the original models, the integration step, or in some cases further literature curation and updating of model content or constraints.
1.i Check mass balances (individual models). Failure: Return to Step 1 (or before). If either the host or the pathogen model result in violation of mass conservation constraints, then the respective model needs to be critically evaluated and debugged, so that the offending reaction(s) is/are identified and removed or adjusted appropriately. The appropriate definition and representation of system boundaries is a simple, yet critical step. Consequences of undefined or inappropriately defined system boundaries will

lead to an ill-formulated model that will likely result in mass balance errors. The cartoon illustration in **Figure 2** highlights the appropriate definition of system boundaries when before and after integration of a host with a pathogen reconstruction. The most direct and common consequence of poorly defined boundaries is an ill-formulated description of the optimization problem with subsequent errors in mass balance, resulting in irrelevant and even non-sensical results.

Dimensionality assessment

Failure: Return to Step 2.i. "Failure" of this step constitutes violation of Equation (3). When merging two (or more reconstructions) there must be a mapping between metabolites that are shared by each of the two models. At minimum there must be at least 1 metabolite that is shared between each model, although in practice there are generally at least 30–40 metabolites that are shared. Once compartment specific identification of shared metabolites has been performed, then the two sets of models can be merged through merging the stoichiometric matrices "row-wise." If $m_h + m_p = m_{hp}$, then there has likely been an error in integration [either through formulation of the problem (Step 1) or implementation of the matrix merge (Step 2.i)]. As noted above, in general, $n_{hp} \approx n_h + n_p$, with the approximation being dependent on whether additional constraints or new objective reactions are added in the integrated network.

Check mass balances (host-pathogen model)

Failure: Return to 2.i. If the integrated host-pathogen model results in violation of mass conservation, but the individual models did not, then there was an error in the model integration (Steps 2.i-2.iii). Evaluation of the boundary exchanges of the pathogen should be the first area of critical evaluation.

Functionality test suite

Failure: Return to 1.iii. Depending on the type of error and the type of functional test, this may be "real" or it may reflect incomplete knowledge (such as an incompletely defined biomass function). Failures in the FTS should be analyzed to determine the source of the limited constraint (the FVA calculations 3.ii can be helpful in tracking this within the network). Once the cause of the failure is identified, it needs to be determined if this is the result of erroneous reaction constraints or a real prediction (i.e., a reaction that is active in the "uninfected" state but is inactive in the infected state). Referral to the primary literature is frequently needed to resolve these issues.

Interdependence test

Failure: Return to Step 1 and 2.ii. The lack of interdependence may require revision of the model(s) (through additional curation and scope expansion) and/or re-assessment of the new constraints and objective functions that were added. For example, in the toy model depicted in **Figure 4Aiii**, further evaluation of the literature may suggest that R15 and/or R16 are active in the pathogen during infections, which would require further evaluation as to how metabolites F and/or X, respectively are made available to the pathogen inside the host cell.

Simulation

Inconsistencies between model predictions and observed experimental results or invalidating predictions should first be assessed in terms of the model and how the specific prediction was made, i.e., identification of the specific pathways leading to the calculated results. If there is no evidence to suggest a model related or numerical error, then there will need to be further perusal of the literature. For example in **Figure 4Aiii**, if there is biochemical or physiologic evidence in the literature suggesting that biochemical transformation carried out by R10 should be active (and able to carry a flux) in the infected state, then there needs to be further evaluation of the literature to determine how metabolite F is taken into the cell, or if there exists an alternative pathway for production of metabolite F within the pathogen (or host). This example also highlights the need for multiple iterative steps that often necessitate re-evaluation of the primary literature. In this case, the pathogen is still able to grow within the host, so there were no errors in Steps 3.i, 3.ii, 3.iii, or 3.iv (assuming that R10 was not contained in the FTS). This example is also illustrative of the need for the multiple checkpoints in the protocol (**Figure 3**) and the necessity of re-evaluating results and possibly revising the model(s) at each step of the integration process.

Current State of the Art and Future Outlook

The systematic procedure described above enables construction of host-pathogen constraint-based models that is applicable to organisms ranging from obligate parasites to multi-cellular pathogens, including viruses, bacteria, and fungi. The methods described above are most directly relevant and applicable to bacterial and fungal organisms. Viruses and parasitic organisms each demonstrate characteristics that may require further considerations, particularly with respect to conditional (e.g., transcription) dependent constraints. Some parasites are multi-cellular organisms that are capable of residing in multiple tissues within a host, thus the challenge by some of these organisms will require the integration of multiple, multi-cellular models. This process will be more involved, but will include the same systematic process. One should recognize the importance of "buffering" compartments and should include them, as they may play an important role in balancing protons, water, phosphate, etc.

Achievement of the steps outlined in **Figure 3** will result in a functional host-pathogen model that should represent a more biologically accurate, quantitative, simulatable description of the interaction between a host and pathogen (**Figure 5**), in turn enabling a more objective, quantitative assessment of the interactions between these cells. Interrogation of these hp models would allow probing pathogen adaptation and carbon source utilization *in vivo* and host manipulation by pathogen. Such models should then be used to answer questions regarding causality during the infection process, condition dependent (or context specific) differences, and ultimately advance diagnosis and treatment related challenges by providing an environment to evaluate and generate hypothesis as well as interpret and analyze data.

The ability to measure and represent data on a genome-scale and the development of constraints based modeling strategies can help explore the complex host-pathogen interaction space (**Figure 5**). While the methods have reached a degree of maturity that enable the application to a wide range of conditions, there still remain many areas that deserve further exploration, including more elegant representation of changes in the environment (e.g., pH changes between different compartments and the associated charge changes that may occur with certain species) as well as more fluid descriptions in the transitions between different growth stages (e.g., rather than static representations for each stage, developing the analog of kinetic models, in which the change from one state to another can be simulated).

The process of host infection is complex and future developments will build upon studies that have, for example, investigated immune responsive signaling pathways such as the Toll-like receptor (Li et al., 2009) as well as the dynamics of pathogen metabolism (Penkler et al., 2015). With continual developments in approaches to expand the scope of reconstructions (Thiele et al., 2009; Lerman et al., 2012) and the development of new methods and approaches for generating genome scale network reconstructions (Overbeek et al., 2005; Henry et al., 2010; Monk et al., 2013), it is anticipated that there will be a dramatic rise in the development of hp models. Ultimately the objective of integrative constraint-based methods is to develop new strategies for treatment of pathogenic infections through novel target identification and new combination therapies for treatment (Trawick and Schilling, 2006; Jamshidi and Palsson, 2007; Karlsson et al., 2011; Chavali et al., 2012).

Constraint-based modeling allows meeting the challenge of complex omic data integration across time and space at multiple levels of hierarchy in the reductionist causal chain to shrink and explore the solution space of host-pathogen interaction. On a genome-scale, multi-cellular level, constraint-based hp modeling has great potential for the prediction of resultant physiologically perturbed cellular states. Implementation across these hierarchical levels of resolution (individual metabolites to mulit-cellular inter-species interactions) at several levels of abstraction will hopefully lead to further elucidation of the metabolic underpinnings of the acute and chronic process of infection, emergent mechanisms of pathogenesis, and therapeutic strategies to counteract such changes.

References

Ahn, S. Y., Jamshidi, N., Mo, M. L., Wu, W., Eraly, S. A., Dnyanmote, A., et al. (2011). Linkage of organic anion transporter-1 to metabolic pathways through integrated "omics"-driven network and functional analysis. *J. Biol. Chem.* 286, 31522–31531. doi: 10.1074/jbc.M111.272534

Barrett, C. L., Price, N. D., and Palsson, B. O. (2006). Network-level analysis of metabolic regulation in the human red blood cell using random sampling and singular value decomposition. *BMC Bioinformatics* 7:132. doi: 10.1186/1471-2105-7-132

Bazzani, S., Hoppe, A., and Holzhütter, H. G. (2012). Network-based assessment of the selectivity of metabolic drug targets in Plasmodium falciparum with respect to human liver metabolism. *BMC Syst. Biol.* 6:118. doi: 10.1186/1752-0509-6-118

Beg, Q. K., Vazquez, A., Ernst, J., de Menezes, M. A., Bar-Joseph, Z., Barabási, A. L., et al. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12663–12668. doi: 10.1073/pnas.0609845104

Beste, D. J., Nöh, K., Niedenführ, S., Mendum, T. A., Hawkins, N. D., Ward, J. L., et al. (2013). 13C-flux spectral analysis of host-pathogen metabolism reveals a mixed diet for intracellular Mycobacterium tuberculosis. *Chem. Biol.* 20, 1012–1021. doi: 10.1016/j.chembiol.2013.06.012

Blazier, A. S., and Papin, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.* 3:299. doi: 10.3389/fphys.2012.00299

Bordbar, A., Feist, A. M., Usaite-Black, R., Woodcock, J., Palsson, B. O., and Famili, I. (2011). A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. *BMC Syst. Biol.* 5:180. doi: 10.1186/1752-0509-5-180

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., and Jamshidi, N. (2010). Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.* 6:422. doi: 10.1038/msb.2010.68

Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* 6:e1000859. doi: 10.1371/journal.pcbi.1000859

Brynildsen, M. P., Winkler, J. A., Spina, C. S., MacDonald, I. C., and Collins, J. J. (2013). Potentiating antibacterial activity by predictably enhancing endogenous microbial ROS production. *Nat. Biotechnol.* 31, 160–165. doi: 10.1038/nbt.2458

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657. doi: 10.1002/bit.10803

Burgard, A. P., Vaidyaraman, S., and Maranas, C. D. (2001). Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol. Prog.* 17, 791–797. doi: 10.1021/bp0100880

Cakir, T., Patil, K. R., Onsan, Z., Ulgen, K. O., Kirdar, B., and Nielsen, J. (2006). Integration of metabolome data with metabolic networks reveals reporter reactions. *Mol. Syst. Biol.* 2, 50. doi: 10.1038/msb4100085

Chang, H. H., Cohen, T., Grad, Y. H., Hanage, W. P., O'Brien, T. F., and Lipsitch, M. (2015). Origin and proliferation of multiple-drug resistance in bacterial pathogens. *Microbiol. Mol. Biol. Rev.* 79, 101–116. doi: 10.1128/MMBR.00039-14

Chang, R. L., Ghamsari, L., Manichaikul, A., Hom, E. F., Balaji, S., Fu, W., et al. (2011). Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* 7, 518. doi: 10.1038/msb.2011.52

Chavali, A. K., D'Auria, K. M., Hewlett, E. L., Pearson, R. D., and Papin, J. A. (2012). A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends Microbiol.* 20, 113–123. doi: 10.1016/j.tim.2011.12.004

Chavali, A. K., Whittemore, J. D., Eddy, J. A., Williams, K. T., and Papin, J. A. (2008). Systems analysis of metabolism in the pathogenic trypanosomatid Leishmania major. *Mol. Syst. Biol.* 4, 177. doi: 10.1038/msb.2008.15

Deatherage Kaiser, B. L., Li, J., Sanford, J. A., Kim, Y. M., Kronewitter, S. R., Jones, M. B., et al. (2013). A multi-omic view of host-pathogen-commensal interplay in -mediated intestinal infection. *PLoS ONE* 8:e67155. doi: 10.1371/journal.pone.0067155

Decker, T., and Lohmann-Matthes, M. L. (1988). A quick and simple method for the quantitation of lactate dehydrogenase release in measurements of cellular cytotoxicity and tumor necrosis factor (TNF) activity. *J. Immunol. Methods* 115, 61–69. doi: 10.1016/0022-1759(88)90310-9

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* 36, D344–D350. doi: 10.1093/nar/gkm791

D'Huys, P. J., Lule, I., Vercammen, D., Anné, J., Van Impe, J. F., and Bernaerts, K. (2012). Genome-scale metabolic flux analysis of Streptomyces lividans growing on a complex medium. *J. Biotechnol.* 161, 1–13. doi: 10.1016/j.jbiotec.2012.04.010

Dräger, A., and Palsson, B. Ø. (2014). Improving collaboration by standardization efforts in systems biology. *Front. Bioeng. Biotechnol.* 2:61. doi: 10.3389/fbioe.2014.00061

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., et al. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1777–1782. doi: 10.1073/pnas.0610772104

Durmuş, S., Çakir, T., Özgür, A., and Guthke, R. (2015). A review on computational systems biology of pathogen-host interactions. *Front. Microbiol.* 6:235. doi: 10.3389/fmicb.2015.00235

Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol.* 7:74. doi: 10.1186/1752-0509-7-74

Edwards, J. S., Ramakrishna, R., and Palsson, B. Ø. (2002). Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol. Bioeng.* 77, 27–36. doi: 10.1002/bit.10047

Ellis, T., Wang, X., and Collins, J. J. (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.* 27, 465–471. doi: 10.1038/nbt.1536

Estrada, E., and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 71:056103. doi: 10.1103/PhysRevE.71.056103

Famili, I., and Palsson, B. O. (2003). The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys. J.* 85, 16–26. doi: 10.1016/S0006-3495(03)74450-6

Fatumo, S., Plaimas, K., Adebiyi, E., and Konig, R. (2011). Comparing metabolic network models based on genomic and automatically inferred enzyme information from Plasmodium and its human host to define drug targets in silico. *Infect. Genet. Evol.* 11, 708–715. doi: 10.1016/j.meegid.2011.04.013

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., and Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* 7, 129–143. doi: 10.1038/nrmicro1949

Feist, A. M., and Palsson, B. O. (2010). The biomass objective function. *Curr. Opin. Microbiol.* 13, 344–349. doi: 10.1016/j.mib.2010.03.003

Fell, D. A., and Small, J. R. (1986). Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* 238, 781–786. doi: 10.1042/bj2380781

Frezza, C., Zheng, L., Folger, O., Rajagopalan, K. N., MacKenzie, E. D., Jerby, L., et al. (2011). Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* 477, 225–228. doi: 10.1038/nature10363

García Saánchez, C. E., and Torres Sáez, R. G. (2014). Comparison and analysis of objective functions in flux balance analysis. *Biotechnol. Prog.* 30, 985–991. doi: 10.1002/btpr.1949

Gawronski, J. D., Wong, S. M., Giannoukos, G., Ward, D. V., and Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16422–16427. doi: 10.1073/pnas.0906627106

Gianchandani, E. P., Oberhardt, M. A., Burgard, A. P., Maranas, C. D., and Papin, J. A. (2008). Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics* 9:43. doi: 10.1186/1471-2105-9-43

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7821–7826. doi: 10.1073/pnas.122653799

Glickman, M. S., Cox, J. S., and Jacobs, W. R. Jr. (2000). A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of Mycobacterium tuberculosis. *Mol. Cell* 5, 717–727. doi: 10.1016/S1097-2765(00)80250-6

Han, J., Antunes, L. C., Finlay, B. B., and Borchers, C. H. (2010). Metabolomics: towards understanding host-microbe interactions. *Future Microbiol.* 5, 153–161. doi: 10.2217/fmb.09.132

Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* 7, 1104–1115. doi: 10.1016/j.celrep.2014.03.070

Henningham, A., Döhrmann, S., Nizet, V., and Cole, J. N. (2015). Mechanisms of group A Streptococcus resistance to reactive oxygen species. *FEMS Microbiol. Rev.* 39, 488–508. doi: 10.1093/femsre/fuu009

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi: 10.1038/nbt.1672

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 26, 1155–1160. doi: 10.1038/nbt1492

Huthmacher, C., Hoppe, A., Bulik, S., and Holzhutter, H. G. (2010). Antimalarial drug targets in Plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC Syst. Biol.* 4:120. doi: 10.1186/1752-0509-4-120

Jamshidi, N., and Palsson, B. Ø. (2007). Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Syst. Biol.* 1:26. doi: 10.1186/1752-0509-1-26

Kafsack, B. F., and Llinás, M. (2010). Eating at the table of another: metabolomics of host-parasite interactions. *Cell Host Microbe* 7, 90–99. doi: 10.1016/j.chom.2010.01.008

Karlsson, F. H., Nookaew, I., Petranovic, D., and Nielsen, J. (2011). Prospects for systems biology and modeling of the gut microbiome. *Trends Biotechnol.* 29, 251–258. doi: 10.1016/j.tibtech.2011.01.009

Khannapho, C., Zhao, H., Bonde, B. K., Kierzek, A. M., Avignone-Rossa, C. A., and Bushell, M. E. (2008). Selection of objective function in genome scale flux balance analysis for process feed development in antibiotic production. *Metab. Eng.* 10, 227–233. doi: 10.1016/j.ymben.2008.06.003

Kim, K., and Weiss, L. M. (2008). Toxoplasma: the next 100years. *Microbes Infect.* 10, 978–984. doi: 10.1016/j.micinf.2008.07.015

Kim, Y. M., Schmidt, B. J., Kidwai, A. S., Jones, M. B., Deatherage Kaiser, B. L., Brewer, H. M., et al. (2013). Salmonella modulates metabolism during growth under conditions that induce expression of virulence genes. *Mol. Biosyst.* 9, 1522–1534. doi: 10.1039/c3mb25598k

Korzeniewski, C., and Callewaert, D. M. (1983). An enzyme-release assay for natural cytotoxicity. *J. Immunol. Methods* 64, 313–320. doi: 10.1016/0022-1759(83)90438-6

Kumar, V. S., and Maranas, C. D. (2009). GrowMatch: an automated method for reconciling *in silico/in vivo* growth predictions. *PLoS Comput. Biol.* 5:e1000308. doi: 10.1371/journal.pcbi.1000308

Kümmel, A., Panke, S., and Heinemann, M. (2006a). Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* 2, 0034. doi: 10.1038/msb4100074

Kümmel, A., Panke, S., and Heinemann, M. (2006b). Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 7:512. doi: 10.1186/1471-2105-7-512

Le Chevalier, F., Cascioferro, A., Majlessi, L., Herrmann, J. L., and Brosch, R. (2014). Mycobacterium tuberculosis evolutionary pathogenesis and its putative impact on drug development. *Future Microbiol.* 9, 969–985. doi: 10.2217/fmb.14.70

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., et al. (2012). In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* 3, 929. doi: 10.1038/ncomms1928

Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305. doi: 10.1038/nrmicro2737

Lewis, N. E., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M. P., Cheng, J. K., et al. (2010). Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nat. Biotechnol.* 28, 1279–1285. doi: 10.1038/nbt.1711

Li, F., Thiele, I., Jamshidi, N., and Palsson, B. Ø. (2009). Identification of potential pathway mediation targets in Toll-like receptor signaling. *PLoS Comput. Biol.* 5:e1000292. doi: 10.1371/annotation/5cc0d918-83b8-44e4-9778-b96a249d4099

Liao, Y. C., Tsai, M. H., Chen, F. C., and Hsiung, C. A. (2012). GEMSiRV: a software platform for GEnome-scale metabolic model simulation, reconstruction and visualization. *Bioinformatics* 28, 1752–1758. doi: 10.1093/bioinformatics/bts267

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10, e1003580. doi: 10.1371/journal.pcbi.1003580

Mahadevan, R., and Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 5, 264–276. doi: 10.1016/j.ymben.2003.09.002

McAdam, P. R., Richardson, E. J., and Fitzgerald, J. R. (2014). High-throughput sequencing for the study of bacterial pathogen biology. *Curr. Opin. Microbiol.* 19, 106–113. doi: 10.1016/j.mib.2014.06.002

Mcconville, M. (2014). Open questions: microbes, metabolism and host-pathogen interactions. *BMC Biol.* 12:18. doi: 10.1186/1741-7007-12-18

Metris, A., George, S., and Baranyi, J. (2012). Modelling osmotic stress by Flux Balance Analysis at the genomic scale. *Int. J. Food Microbiol.* 152, 123–128. doi: 10.1016/j.ijfoodmicro.2011.06.016

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., and Shlomi, T. (2012). Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Natl. Acad. Sci. U.S.A.* 109, 339–344. doi: 10.1073/pnas.1100358109

Mo, M. L., Jamshidi, N., and Palsson, B. Ø. (2007). A genome-scale, constraint-based approach to systems biology of human metabolism. *Mol. Biosyst.* 3, 598–603. doi: 10.1039/b705597h

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., et al. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20338–20343. doi: 10.1073/pnas.1307797110

Nairz, M., Ferring-Appel, D., Casarrubea, D., Sonnweber, T., Viatte, L., Schroll, A., et al. (2015). Iron regulatory proteins mediate host resistance to salmonella infection. *Cell Host Microbe* 18, 254–261. doi: 10.1016/j.chom.2015.06.017

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614

Osterlund, T., Nookaew, I., and Nielsen, J. (2012). Fifteen years of large scale metabolic modeling of yeast: developments and impacts. *Biotechnol. Adv.* 30, 979–988. doi: 10.1016/j.biotechadv.2011.07.021

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702. doi: 10.1093/nar/gki866

Pacchiarotta, T., Deelder, A. M., and Mayboroda, O. A. (2012). Metabolomic investigations of human infections. *Bioanalysis* 4, 919–925. doi: 10.4155/bio.12.61

Palsson, B. (2015). *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge: Cambridge University Press.

Pan, X., Tamilselvam, B., Hansen, E. J., and Daefler, S. (2010). Modulation of iron homeostasis in macrophages by bacterial intracellular pathogens. *BMC Microbiol.* 10:64. doi: 10.1186/1471-2180-10-64

Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol. Bioeng.* 26, 174–187. doi: 10.1002/bit.260260210

Penkler, G., Du Toit, F., Adams, W., Rautenbach, M., Palm, D. C., Van Niekerk, D. D., et al. (2015). Construction and validation of a detailed kinetic model of glycolysis in Plasmodium falciparum. *FEBS J.* 282, 1481–1511. doi: 10.1111/febs.13237

Phalakornkule, C., Lee, S., Zhu, T., Koepsel, R., Ataai, M. M., Grossmann, I. E., et al. (2001). A MILP-based flux alternative generation and NMR experimental design strategy for metabolic engineering. *Metab. Eng.* 3, 124–137. doi: 10.1006/mben.2000.0165

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 14, 2367–2376. doi: 10.1101/gr.2872004

Pornputtapong, N., Nookaew, I., and Nielsen, J. (2015). Human metabolic atlas: an online resource for human metabolism. *Database (Oxford)* 2015, bav068. doi: 10.1093/database/bav068

Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897. doi: 10.1038/nrmicro1023

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., et al. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst. Biol.* 4:114. doi: 10.1186/1752-0509-4-114

Raghunathan, A., Reed, J., Shin, S., Palsson, B., and Daefler, S. (2009). Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. *BMC Syst. Biol.* 3:38. doi: 10.1186/1752-0509-3-38

Raghunathan, A., Shin, S., and Daefler, S. (2010). Systems approach to investigating host-pathogen interactions in infections with the biothreat agent Francisella. Constraints-based model of Francisella tularensis. *BMC Syst Biol* 4:118. doi: 10.1186/1752-0509-4-118

Ratledge, C., and Dover, L. G. (2000). Iron metabolism in pathogenic bacteria. *Annu. Rev. Microbiol.* 54, 881–941. doi: 10.1146/annurev.micro.54.1.881

Reed, J. L., and Palsson, B. Ø. (2003). Thirteen years of building constraint-based in silico models of *Escherichia coli*. *J. Bacteriol.* 185, 2692–2699. doi: 10.1128/JB.185.9.2692-2699.2003

Robaina Estévez, S., and Nikoloski, Z. (2014). Generalized framework for context-specific metabolic model extraction methods. *Front. Plant Sci.* 5:491. doi: 10.3389/fpls.2014.00491

Rodriguez, G. M., Voskuil, M. I., Gold, B., Schoolnik, G. K., and Smith, I. (2002). ideR, An essential gene in mycobacterium tuberculosis: role of IdeR in iron-dependent gene expression, iron metabolism, and oxidative stress response. *Infect. Immun.* 70, 3371–3381. doi: 10.1128/IAI.70.7.3371-3381.2002

Sadhukhan, P. P., and Raghunathan, A. (2014). Investigating host-pathogen behavior and their interaction using genome-scale metabolic network models. *Methods Mol. Biol.* 1184, 523–562. doi: 10.1007/978-1-4939-1115-8_29

Saha, R., Suthers, P. F., and Maranas, C. D. (2011). Zea mays iRS1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE* 6:e21784. doi: 10.1371/journal.pone.0021784

Sauro, H. M., and Ingalls, B. (2004). Conservation analysis in biochemical networks: computational issues for software writers. *Biophys. Chem.* 109, 1–15. doi: 10.1016/j.bpc.2003.08.009

Savinell, J. M., and Palsson, B. O. (1992). Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J. Theor. Biol.* 155, 201–214. doi: 10.1016/S0022-5193(05)80595-8

Schellenberger, J., and Palsson, B. Ø. (2009). Use of randomized sampling for analysis of metabolic networks. *J. Biol. Chem.* 284, 5457–5461. doi: 10.1074/jbc.R800048200

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., et al. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* 6, 1290–1307. doi: 10.1038/nprot.2011.308

Schoen, C., Kischkies, L., Elias, J., and Ampattu, B. J. (2014). Metabolism and virulence in Neisseria meningitidis. *Front. Cell. Infect. Microbiol.* 4:114. doi: 10.3389/fcimb.2014.00114

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* 3, 119. doi: 10.1038/msb4100162

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science* 336, 601–604. doi: 10.1126/science.1216882

Seaver, S. M., Henry, C. S., and Hanson, A. D. (2012). Frontiers in metabolic reconstruction and modeling of plant genomes. *J. Exp. Bot.* 63, 2247–2258. doi: 10.1093/jxb/err371

Shoaie, S., and Nielsen, J. (2014). Elucidating the interactions between the human gut microbiota and its host through metabolic modeling. *Front. Genet.* 5:86. doi: 10.3389/fgene.2014.00086

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., et al. (2012). Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336, 1157–1160. doi: 10.1126/science.1217405

Stavrinides, J., McCann, H. C., and Guttman, D. S. (2008). Host-pathogen interplay and the evolution of bacterial effectors. *Cell. Microbiol.* 10, 285–292. doi: 10.1111/j.1462-5822.2007.01078.x

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., et al. (2007). Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* 3, 92. doi: 10.1038/msb4100131

Szekely, P., Sheftel, H., Mayo, A., and Alon, U. (2013). Evolutionary tradeoffs between economy and effectiveness in biological homeostasis systems. *PLoS Comput. Biol.* 9:e1003163. doi: 10.1371/journal.pcbi.1003163

Takayama, K., Wang, C., and Besra, G. S. (2005). Pathway to synthesis and processing of mycolic acids in Mycobacterium tuberculosis. *Clin. Microbiol. Rev.* 18, 81–101. doi: 10.1128/CMR.18.1.81-101.2005

Thiele, I., Jamshidi, N., Fleming, R. M., and Palsson, B. Ø. (2009). Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational

machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS Comput. Biol.* 5:e1000312. doi: 10.1371/journal.pcbi.1000312

Thiele, I., and Palsson, B. Ø. (2010). Reconstruction annotation jamborees: a community approach to systems biology. *Mol. Syst. Biol.* 6, 361. doi: 10.1038/msb.2010.15

Thomas, A., Rahmanian, S., Bordbar, A., Palsson, B. Ø., and Jamshidi, N. (2014). Network reconstruction of platelet metabolism identifies metabolic signature for aspirin resistance. *Sci. Rep.* 4, 3925. doi: 10.1038/srep03925

Trawick, J. D., and Schilling, C. H. (2006). Use of constraint-based modeling for the prediction and validation of antimicrobial targets. *Biochem. Pharmacol.* 71, 1026–1035. doi: 10.1016/j.bcp.2005.10.049

Tymoshenko, S., Oppenheim, R. D., Agren, R., Nielsen, J., Soldati-Favre, D., and Hatzimanikatis, V. (2015). Metabolic needs and capabilities of toxoplasma gondii through combined computational and experimental analysis. *PLoS Comput. Biol.* 11:e1004261. doi: 10.1371/journal.pcbi.1004261

Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., et al. (2015). Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell Rep.* 11, 921–933. doi: 10.1016/j.celrep.2015.04.010

Varma, A., Boesch, B. W., and Palsson, B. Ø. (1993). Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* 59, 2465–2473.

Virchow, R. (1958). *Cellular Pathology.* Stanford, CA: Stanford Univesity Press.

Wang, Y., Eddy, J. A., and Price, N. D. (2012). Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* 6:153. doi: 10.1186/1752-0509-6-153

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Model.* 28, 31–36. doi: 10.1021/ci00057a005

Weiss, G., and Schaible, U. E. (2015). Macrophage defense mechanisms against intracellular bacteria. *Immunol. Rev.* 264, 182–203. doi: 10.1111/imr.12266

Yao, J., and Rock, C. O. (2015). How bacterial pathogens eat host lipids: implications for the development of fatty acid synthesis therapeutics. *J. Biol. Chem.* 290, 5940–5946. doi: 10.1074/jbc.R114.636241

Zakrzewski, P., Medema, M. H., Gevorgyan, A., Kierzek, A. M., Breitling, R., and Takano, E. (2012). MultiMetEval: comparative and multi-objective analysis of genome-scale metabolic models. *PLoS ONE* 7:e51511. doi: 10.1371/journal.pone.0051511

Zomorrodi, A. R., Islam, M. M., and Maranas, C. D. (2014). d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth. Biol.* 3, 247–257. doi: 10.1021/sb4001307

Zomorrodi, A. R., and Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* 8:e1002363. doi: 10.1371/journal.pcbi.1002363

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 30 June 2015
doi: 10.3389/fmicb.2015.00625

Host-pathogen interactions between the human innate immune system and *Candida albicans*—understanding and modeling defense and evasion strategies

*Sybille Dühring[1], Sebastian Germerodt[1], Christine Skerka[2], Peter F. Zipfel[2,3], Thomas Dandekar[4] and Stefan Schuster[1]**

[1] Department of Bioinformatics, Friedrich-Schiller-University Jena, Jena, Germany, [2] Department of Infection Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany, [3] Friedrich-Schiller-University Jena, Jena, Germany, [4] Department of Bioinformatics, Biozentrum, Universitaet Wuerzburg, Wuerzburg, Germany

OPEN ACCESS

Edited by:
Saliha Durmus, Gebze Technical University, Turkey

Reviewed by:
Mihai Netea, Radboud University Nijmegen Medical Center, Netherlands
Mehmet Mete Altintas, Rush University, USA

***Correspondence:**
Stefan Schuster, Department of Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany
stefan.schu@uni-jena.de

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *16 March 2015*
Accepted: *08 June 2015*
Published: *30 June 2015*

Citation:
Dühring S, Germerodt S, Skerka C, Zipfel PF, Dandekar T and Schuster S (2015) Host-pathogen interactions between the human innate immune system and Candida albicans—understanding and modeling defense and evasion strategies. Front. Microbiol. 6:625. doi: 10.3389/fmicb.2015.00625

The diploid, polymorphic yeast *Candida albicans* is one of the most important human pathogenic fungi. *C. albicans* can grow, proliferate and coexist as a commensal on or within the human host for a long time. However, alterations in the host environment can render *C. albicans* virulent. In this review, we describe the immunological cross-talk between *C. albicans* and the human innate immune system. We give an overview in form of pairs of human defense strategies including immunological mechanisms as well as general stressors such as nutrient limitation, pH, fever etc. and the corresponding fungal response and evasion mechanisms. Furthermore, Computational Systems Biology approaches to model and investigate these complex interactions are highlighted with a special focus on game-theoretical methods and agent-based models. An outlook on interesting questions to be tackled by Systems Biology regarding entangled defense and evasion mechanisms is given.

Keywords: *Candida albicans*, human immune system, host-pathogen interaction, computational systems biology, defense and evasion strategies, immunological cross-talk

1. Introduction

The diploid, polymorphic yeast *Candida albicans* (Wilson et al., 2009; Kwak et al., 2014; Mech et al., 2014) is one of the most important human pathogenic fungi (Lu et al., 2014; Vylkova and Lorenz, 2014; Whittington et al., 2014). This opportunistic ubiquitous fungus (Faro-Trindade and Brown, 2009; Zipfel et al., 2011; Bain et al., 2012) usually resides as a commensal on the skin and mucosal surfaces of 30 to 70 % of the human population (Cheng et al., 2012; Jacobsen et al., 2012; Quintin et al., 2014). As part of the normal human microbiota in the gastrointestinal, oropharyngeal or urogenital tract (Moyes and Naglik, 2011; Luo et al., 2013; Wellington et al., 2014) *C. albicans* can grow, proliferate and coexist within the human host for a long time (Yan et al., 2013) without causing any symptoms of disease (Moyes and Naglik, 2011; Gow et al., 2012; Mayer et al., 2013).

C. albicans is highly specialized for the life on or within the human host (Wilson et al., 2009). The homeostasis between *C. albicans* and the human host is kept by the human immune system (Luo et al., 2013; Yan et al., 2013; Vylkova and Lorenz, 2014) and the normal bacterial flora on mucosal surfaces and epithelial layers (Mayer et al., 2013; Yan et al., 2013; Mech et al., 2014). However, alterations in the host environment can render commensal factors into virulence attributes once the conditions favor pathogenicity (Moyes and Naglik, 2011; Bain et al., 2012; Whittington et al., 2014). Thus, there is a subtle balance between the commensal- and the pathogenic state of *C. albicans*. This is testified by a large number of defense mechanisms of the immune system and evasion mechanisms of *C. albicans*. Furthermore, there is evidence for probiotic action of *C. albicans*, for instance regarding protection of the vaginal flora (Martin et al., 1999). To understand this complex interplay, Systems Biology approaches have shown to be very instrumental (Hummert et al., 2010; Mech et al., 2014; Tierney et al., 2014). Here we provide a systematic overview of the host-pathogen interactions to promote this endeavor.

There are two major types of *C. albicans* infections in humans (Filler, 2013; Luo et al., 2013; Mayer et al., 2013). Superficial mucosal diseases like vaginal or oral candidiasis are extremely common (Cheng et al., 2012; Filler, 2013; Wellington et al., 2014). It is estimated that 75 % of all women worldwide suffer from vulvovaginal candidiasis at least once in their life and 40 to 50 % experience recurrent infections (Wilson et al., 2009; Mayer et al., 2013). Startlingly vaginal infections often occur without any sign of immune defect (Jacobsen et al., 2012). While *Candida*-associated denture stomatitis is caused in elderly and edentulous individuals (Moyes and Naglik, 2011; Mayer et al., 2013) oral and oesophageal candidiasis is particularly common in HIV-positive individuals (Wilson et al., 2009; Moyes and Naglik, 2011; Jacobsen et al., 2012). Severe mucosal diseases and life-threatening systemic infections arise in immunocompromised individuals (Yan et al., 2013; Mech et al., 2014; Wellington et al., 2014). These invasive infections of the bloodstream and virtually every organ of the human body (Wilson et al., 2009; Mayer et al., 2013; Vylkova and Lorenz, 2014) are associated with a severe morbidity (Faro-Trindade and Brown, 2009; Zipfel et al., 2011; Cheng et al., 2012; Luo et al., 2013), an unexeptably high mortality (Zipfel et al., 2011; Filler, 2013; Yan et al., 2013) and high healthcare costs (Yan et al., 2013; Vialas et al., 2014). As disseminated hematogenous candidiasis is the 3rd to 4th most common nosocomial bloodstream infection (Faro-Trindade and Brown, 2009; Wilson et al., 2009; Vylkova and Lorenz, 2014) *C. albicans* is medically as important as many mainstream bacterial infections including *Enterococci* like *Escherichia coli* and *Pseudomonas* spp. (Moyes and Naglik, 2011; Zipfel et al., 2011).

It is of particular significance that the human immune system is able to discriminate between the commensal colonization and the pathogenic invasion phase of *C. albicans* (Cheng et al., 2012; Gow et al., 2012). A robust immune response is therefore required to protect the host against *Candida* infection (d'Enfert, 2009; Jacobsen et al., 2012). This immune response can be divided into physical barriers and immune-barriers of the mucosa (Luo et al., 2013; Yan et al., 2013). The complexity of possible host-pathogen-interactions is high, as *C. albicans* is likely to encounter different components of the human immune system (Jacobsen et al., 2012). *C. albicans* has developed a large number of strategies to evade or undermine the antimicrobial defense responses of the host immune system (Collette and Lorenz, 2011; Zipfel et al., 2011; Lopez, 2013; Luo et al., 2013). These strategies may allow the fungus to control the host immune attack, to cross tissue barriers and to disseminate in the human body (Zipfel et al., 2011; Jacobsen et al., 2012; Luo et al., 2013).

In this review, we describe the immunological cross-talk between *C. albicans* and the human immune system. We follow the trail of infection: starting from phenotypic adaptations and morphogenesis, *Candida* encounters stress by the host and its infected tissue environment including nutrient limitation, temperature and pH stress. Furthermore, the host immune system responds with the innate immune attack as the immediately acting primary line of defense against systemic fungal infections (Cheng et al., 2012; Lopez, 2013). That line of host defense mainly relies on humoral complement actions, antimicrobial peptides and the cellular response mediated by phagocytes, especially by neutrophils and macrophages (Bain et al., 2012; Cheng et al., 2012; Jiménez-López and Lorenz, 2013).

Because of the complex immune response we here focus on the innate immune system. The adaptive immune system contributes with many additional mechanisms (Curtis and Way, 2009; Korn et al., 2009; Hamad, 2012) which are beyond the scope of this review. For a review on the crosstalk between innate and adaptive immune-response and the role of dendritic cells see Hamad (2012). Here, we start to give an overview of the crosstalk of human defense mechanisms and the corresponding fungal evasion mechanisms. As the total amount of such mechanisms is enormous, the list cannot be exhaustive.

A number of different Systems Biology approaches exist to model and simulate host pathogen interactions, e.g., Boolean modeling (Naseem et al., 2012; Schlatter et al., 2012) and reverse engineering (Tierney et al., 2012). However, no matter which strategy is chosen, there is always a game of life and death involved and hence game theoretical approaches and agent-based modeling are particularly powerful and thus reviewed here. These approaches are useful to depict the highly complex and dynamic host-pathogen interactions and can help to gain further insights into the underlying processes of *C. albicans* infections.

2. Host Defense and Corresponding Fungal Evasion Strategies

As a commensal as well as an invading pathogen *C. albicans* faces stressors of the host environment. Those stress sources include changes in nutrient availability, pH, osmolarity, temperature, or attack by the cells of the immune system (Wilson et al., 2009). *C. albicans* has a robust stress response mediated by humoral components as well as rapid alterations in gene expression of stress-responsive regulatory pathways which allow *C. albicans* to respond to changes of environmental stimuli (Wilson et al., 2009; Mayer et al., 2013).

2.1. Stress Induced by the Human Host and its Environment

As a commensal *C. albicans* competes with all the probiotic microorganisms of the host's microflora for nutrients (Brunke and Hube, 2013; Whittington et al., 2014). Even though the gut is relatively rich in nutrients, those nutrients are quickly absorbed by the microbial flora and the epithelial cells (Whittington et al., 2014). In other host niches nutrients are limited by the host and usually not available to pathogens (see **Table 1**). However, *C. albicans* is metabolically flexible and uses nutrient acquisition mechanisms such as sequestration of iron and zinc to survive and grow in the many different and changing host niches such as the gastrointestinal, oropharyngeal or urogenital tract (Brunke and Hube, 2013; Whittington et al., 2014).

One of those mechanisms is the release of secreted aspartic proteases (Saps) by *C. albicans*. Saps can destroy host tissue and liberate oligopeptides and amino acids. These liberated carbon sources are then taken up by *C. albicans* via oligopeptide and amino acid transporters (Brunke and Hube, 2013). Adding to the metabolic flexibility *C. albicans* has no known auxotrophies and can metabolize a broad range of sugars and all amino acids (Brunke and Hube, 2013; Lopez, 2013).

When facing nutrient starvation or phagocytosis *C. albicans* shows responses that are similar in the two cases, switching from the glycolytic pathway to the glyoxylate cycle and gluconeogenesis (Faro-Trindade and Brown, 2009; Lopez, 2013; Vylkova and Lorenz, 2014) which is absent from humans. This metabolic shift enables *C. albicans* to metabolize alternative, less favored carbon sources (De Figueiredo et al., 2008; Faro-Trindade and Brown, 2009). Next to amino acids and lipids, lactate produced by tissues and bacteria in the gut serves as one potential carbon source (Lopez, 2013). The growth on alternative carbon sources can cause substantial changes in the cell wall of *C. albicans* even when the morphology of the cell is otherwise unaltered (Gow et al., 2012). This influences the recognition by phagocytes as well as drug- and stress-resistance of the cell (Vylkova and Lorenz, 2014). *C. albicans* cells grown on lactate have been shown to be more resistant to osmotic, envelope and antifungal stresses and to be more adherent. They even elicit lower levels of proinflammatory cytokines from monocytic cells and once phagocytosed are more harmful to macrophages than *C. albicans* cells grown on glucose. The exposure to non-preferred carbon sources therefore benefits *C. albicans*, especially in their interactions with macrophages (Lopez, 2013). This fact may be helpful for metabolic therapy strategies, as supplementing certain nutrients in the infection locus may render *C. albicans* more vulnerable to the host defense.

Another important defense mechanism in the host's "nutritional immunity" is the active sequestration of metals (Brunke and Hube, 2013). The most important micro nutrients that are prerequisite for *C. albicans* infection are iron, zinc, manganese and copper (Brunke and Hube, 2013; Mayer et al., 2013). Both the pathogenic fungus and its host have evolved mechanisms to acquire and restrict access to these metals (Mayer et al., 2013).

The human host is severely restricting the availability of iron to pathogens by keeping the iron levels of the blood and the tissue environment low (Brunke and Hube, 2013). This is achieved by storing iron in iron-binding proteins like ferritin, lactoferrin, hemoglobin and transferrin which are usually not accessible to pathogenic microbes (Faro-Trindade and Brown, 2009; Jacobsen et al., 2012). *C. albicans* on the other side has developed a plethora of iron acquisition systems (Brunke and Hube, 2013) including a reductive system, a siderophore uptake system and a heme-iron uptake system (Mayer et al., 2013). *C. albicans* can utilize its siderophore uptake system via Sit1/Arn1 (siderophore iron transport 1) to steal iron from siderophores produced by other microorganisms without producing its own siderophores (Brunke and Hube, 2013; Mayer et al., 2013). *C. albicans* can further bind host ferritin with the hyphae-associated adhesion and invasion protein Als3 (agglutinin-like sequence 3) (Jacobsen et al., 2012; Brunke and Hube, 2013; Mayer et al., 2013). The reductive system, with its large gene families of reductases, oxidases and iron permeases (Brunke and Hube, 2013), then

TABLE 1 | Pairs of defense and evasion strategies—adapting to the host environment.

Human host	*C. albicans*
Limiting nutrient availability to pathogens	Release of secreted aspartic proteases (Saps) to liberate oligopeptides and amino acids from tissues
Nutrient starvation e.g., in phagocytes	Switching from the glycolytic pathway to the glyoxylate cycle and gluconeogenesis to metabolize alternative carbon sources
Active sequestration of iron	Iron acquisition through: a reductive system, a siderophore uptake system and a heme-iron uptake system
Active sequestration of zinc	Zinc acquisition via a zincophore system
Inducing pH-stress	Sense and adapt to environmental pH (Pra1); modulate extracellular pH by actively alkalizing the surrounding environment
Inducing thermal stress like fever	Heat shock response mediated by heat shock proteins and trehalose accumulation
Inducing osmotic stress	Outer cell wall structure as protection from osmotic pressure; intracellular accumulation of glycerol to counteract the loss of water

mediates the iron acquisition from host ferritin, transferrin or if available free iron from the environment. *C. albicans* can also use iron from host hemoglobin and hemoproteins by first expressing haemolysins that disrupt red blood cells (Brunke and Hube, 2013; Mayer et al., 2013). Subsequently the iron acquisition is mediated by the heme-receptor gene family members RBT5, RBT51, CSA1, CSA2, and PGA7 (RBT6) (Mayer et al., 2013).

Zinc as a central cofactor for many proteins is an abundant metal in most living organisms (Brunke and Hube, 2013). The sequestration of zinc is therefore a potent antifungal mechanism of the host during infections and is mediated by calprotectin (Faro-Trindade and Brown, 2009; Brunke and Hube, 2013). *C. albicans* can acquire zinc via a "zincophore" system using pH-regulated antigen 1 (Pra1) (Brunke and Hube, 2013). Secreted Pra1 acts as a zincophor, similar to iron-carrying siderophores (Brunke and Hube, 2013), binds extracellular zinc and reassociates with the fungal cell (Mayer et al., 2013). This reassociation is mediated by the zinc transporter Zrt1 (Mayer et al., 2013).

Though copper and manganese are essential for fungal growth, the mechanisms by which *C. albicans* acquires them are less well understood. There is a putative manganese transporter, Ccc1, and a putative copper transporter, Ctr1, but their roles in *C. albicans* virulence have not yet been determined (Mayer et al., 2013).

Next to nutritional stress, pH-stress is of fundamental importance to *C. albicans*. Depending on the host niche *C. albicans* encounters many different pH levels. While the pH of human blood and tissues is slightly alkaline, the pH of the digestive tract ranges from very acidic to more alkaline. pH-stress can also occur in the urogenital tract as well as in the phagolysosome, where pH is very acidic, once *C. albicans* is phagocytosed by cells of the innate immune system. However, *C. albicans* is able to adapt to significant changes in its surrounding pH. The two *C. albicans* cell wall proteins Phr1 (pH responsive 1) (required for systemic infections) and Phr2 (essential for infections of the vagina) are important for adaptation to changing pH. Astonishingly *C. albicans* is not only able to sense and adapt to environmental pH but also to modulate extracellular pH by actively alkalizing its surrounding environment (Mayer et al., 2013). *C. albicans* can release ammonia derived from amino acid degradation to raise extracellular pH (Lopez, 2013). This is of special importance after phagocytosis as it promotes the neutralization of the phagosomal pH, inducing hyphal morphogenesis and thereby fosters the escape of the pathogen from macrophages (Vylkova and Lorenz, 2014).

Thermal stress like fever and cold leads to a heat shock response mediated by heat shock proteins and trehalose accumulation in *C. albicans*. These heat shock proteins and trehalose act as "molecular- and chemical- chaperons" by preventing deleterious protein unfolding and aggregation (Mayer et al., 2013).

The outer layer of *C. albicans*' cell wall not only defines the cell shape and provides an efficient barrier against immune reactions but also protects the fungus from osmotic pressure (Luo et al., 2013). A further osmotic stress response is the intracellular accumulation of glycerol to counteract the loss of water due to an outward-directed chemical gradient. The glycerol biosynthesis is mediated by the glycerol 3-phosphatase (Gpp1) and the glycerol 3-phosphate dehydrogenase 2 (Gpd2) (Mayer et al., 2013).

2.2. The Human Innate Immun System

Zipfel et al. (2011) started a list of immune evasion and tissue invasion mechanisms including complement evasion, evasion of cellular response and tissue invasion mechanisms by *C. albicans* which we include and further augment in **Table 2**. Several of the mechanisms listed in **Table 2** have been modeled by Systems Biology approaches (see Section 3).

C. albicans can switch readily between yeast, hyphal and pseudohyphal growth and back (Brunke and Hube, 2013; Kwak et al., 2014; Lu et al., 2014). Both the yeast and hyphal forms of the fungus are required for biofilm formation as well as full virulence (Lopez, 2013; Yan et al., 2013). Mutants locked in one morphology are avirulent and show a significantly reduced growth performance in biofilm formation (Baillie and Douglas, 1999; Lopez, 2013; Yan et al., 2013).

Biofilms are three-dimensional microbial communities in an extracellular matrix adhering to mucosal or artificial surfaces (Ganguly and Mitchell, 2011) for example biomaterials used for implants like stents and catheters. While the *C. albicans* biofilms on abiotic surfaces consist of yeast and filamentous cells of the fungus (Baillie and Douglas, 1999; Ramage et al., 2006; Ganguly and Mitchell, 2011) the *in vivo C. albicans* biofilms are polymicrobial with an extracellular matrix layer that contains host immune cells (Ganguly and Mitchell, 2011). *C. albicans* biofilms protect the pathogen from host immune attacks and antifungal drugs (Baillie and Douglas, 1998; Seneviratne et al., 2008; Yan et al., 2013). Especially *C. albicans* abiotic surface biofilms are associated with increased drug resistance (Baillie and Douglas, 1998; Ganguly and Mitchell, 2011). This antifungal resistance increases with biofilm maturation (Chandra et al., 2001). There are indications that *C. albicans* biofilms are even resistant to killing by neutrophils (Ganguly and Mitchell, 2011; Mayer et al., 2013) and do not trigger the production of reactive oxygen species (ROS) (Mayer et al., 2013). Reviews on the regulatory control of *C. albicans* within biofilms can be found in Nobile and Mitchell (2006) and Finkel and Mitchell (2011). For a review on *C. albicans* biofilms on mucosal surfaces see Ganguly and Mitchell (2011).

After invading host tissues *C. albicans* encounters an early defense line: the innate immune system. The innate immune system maintains host homeostasis by recognizing and cleaning modified or damaged host cells. It directly attacks and limits the growth of invading microbes without inflammatory reactions. This defense is mediated through three major effector mechanisms: antimicrobial peptides, the complement system and immune cells that recognize and respond to foreign microbes (Zipfel et al., 2011).

The initial interaction of *C. albicans* with the human immune system is with epithelial cells of the mucosa (Moyes and Naglik, 2011; Luo et al., 2013) that act as physical barriers. The fungus is able to invade the human host tissues via two routes: induced endocytosis and active penetration (Mech et al., 2014). The passive uptake is a host driven process, mediated by *C. albicans*

TABLE 2 | Pairs of defense and evasion strategies—*C. albicans* and the human innate immune system.

Human host	*C. albicans*
EPITHELIAL RESPONSE	
Physical barrier	Active penetration by thigmotropism, elongating hyphae and production of lytic enzymes; induction of endocytosis; degradation of extracellular matrix component by recruiting human plasminogen to the yeast surface and secretion of lytic enzymes
Chemical barrier in form of secreted antimicrobial peptides and degradative enzymes	Respond to β-defensin activity via the high-osmolarity glycerol (HOG) pathway; secretion of Sap9 and a Msb2 fragment
The host uses *C. albicans'* pumps to get antimicrobial peptides into the pathogen	Uses multi-drug resistance pumps such as Flu1 to transport antimicrobial peptides out of the pathogen
COMPLEMENT RESPONSE	
Complement systems barrier	Acquiring human complement regulators to the cell surface; secretion of complement inhibitors to block C3 complement activation; production of proteases (Saps) to degrade host complement proteins
CELLULAR RESPONSE	
PRRs recognition barrier via dectin-1, dectin-2, etc.	Surface mannans shield β-glucan from recognition by dectin-1 to avoiding phagocytosis; release of soluble decoys to evade host immune responses
Barrier in form of pro- and anti-inflammatory cytokines and chemokines production	Inhibition of proinflammatory IL-17 production by altering the host tryptophan metabolism; induction of anti-inflammatory cytokine release
Inhibition of *C. albicans* yeast-to-hyphal transition by neutrophils	No known evasion mechanism
Cellular ET formation by neutrophils and macrophages	No known evasion mechanism
Phagocytosis	Biofilm formation; inhibition of phagolysosome formation; neutralization of phagosomal pH inside macrophages; induction of hyphal morphogenesis and escape from the immune cell in macrophages and natural killer cells; pyroptosis / macrophage cell death
Oxidative and nitrosative stress induced by neutrophils and macrophages	Inhibition of ROS generation by macrophages through an unknown mechanism; secretion of Sod enzymes, catalases, glutathione peroxidases and thioredoxin to detoxify extracellular ROS; accumulation of trehalose against oxidative stress; production of intracellular flavohemoglobin enzymes against nitrosative stress; biofilm formation

surface proteins Als3 and Hgc1 (hypha-specific expression and relatedness to G1 cyclins 1) which bind to epithelial cell E-cadherin (Wilson et al., 2009). Active penetration on the other hand does not rely on the host but exclusively on fungal attributes including physical pressure applied by the advancing hyphal tip, thigmotropism and the secretion of extracellular hydrolases like Saps, class B phospholipase (Plb) and lipase (Lip) families (Wilson et al., 2009; Mayer et al., 2013).

Epithelial cells not only provide a physical barrier but also have an active, integral role in mucosal protection against *C. albicans* by discriminating between the commensal and pathogenic form of the fungus. Next to the NF-κB pathway, Moyes and Naglik (2011) identified the MAPK signaling as an important mechanism in the epithelial cell responses to *Candida* infections. The presence of *C. albicans* yeast or hyphae triggers the NF-κB signaling and an early response of the MAPK activation through ERK1/2 and JNK signaling which induces the c-Jun activity. When a sufficient fungal hyphal burden is present and the threshold level of activation is reached, a second prolonged, late response is induced, activating MAPK regulation via the MAPK phosphatase MKP1 through ERK1/2 and p38 signaling. This in turn induces c-Fos activity resulting in the production of cytokines with a proinflammatory profile like interleukin 1α/β (IL-1α/β), IL-6, G-CSF, GM-CSF, and TNF-α as well as the chemokines RANTES, IL-8, and CCL20 (Steele and Fidel, 2002; Moyes and Naglik, 2011; Cheng et al., 2012). Chin et al. (2014) showed that the post-infection regulation of cytokines for IL-2, IL-6, TNF-α, TNF-β are organ-specific (i.e., kidney, spleen, brain). The secretion of proinflammatory molecules results in the recruitment, differentiation, and activation of various immune cells (Moyes and Naglik, 2011; Cheng et al., 2012). Especially important for an early immune response of mucosal surfaces to *C. albicans* infection is IL-22. This cytokine is produced by innate and adaptive immune cells (De Luca et al., 2010; Zenewicz and Flavell, 2011). A heterodimeric receptor consisting of IL-22R and IL-10Rb recognizes IL-22 (Eyerich et al., 2011; Sonnenberg et al., 2011; Zenewicz and Flavell, 2011). As the expression of IL-22R is mainly confined to epithelial cells, the signaling is specific to tissues (Eyerich et al., 2011; Zenewicz and Flavell, 2011). IL-22 has both pro- and anti-inflammatory functions (De Luca et al., 2010) and stimulates the proliferation (Kagami et al., 2010; Zenewicz and Flavell, 2011) and together with IL-17 the production of antimicrobial peptides by epithelial cells (De Luca et al., 2010; Kagami et al., 2010; Eyerich et al., 2011). The IL-23/IL-22 axis controls the initial fungal growth and tissue homeostasis (De Luca et al., 2010; Zelante et al., 2011). The combinatorial secretion of IL-22 and TNF-α by Th22 cells increases the induction and secretion of the complement factors C1r and C1s, antimicrobial chemokines and antimicrobial peptides (Eyerich et al., 2011).

2.2.1. Antimicrobial Peptides

Two important groups of antimicrobial peptides are α- and β-defensins. The α-defensins group consists of four cationic peptides, HNP1 to HNP4, that are found in the azurophilic

granules of human neutrophils. The group of β-defensins is primarily expressed by epithelial cells and includes human β-defensins 2 and 3 (hBD-2 and hBD-3), that have significant antifungal activity. They can be induced by a variety of agents, including TLR agonists, as well as monocyte- and macrophage-derived factors, such as IL-1 (Faro-Trindade and Brown, 2009). The two groups of α- and β-defensins can be distinguished based on their arrangement of disulfide linkages (Faro-Trindade and Brown, 2009; Yan et al., 2013). Both defensin groups target *C. albicans* cell membranes and cause nonlytic permeabilization and release of cellular ATP (Faro-Trindade and Brown, 2009). The damage imposed on *C. albicans* by hBD-2 and hBD-3 shares similarities with that caused by osmotic and oxidative stress. *C. albicans* in turn can respond to these hBD-2 and hBD-3 injuries via the high-osmolarity glycerol (HOG) pathway and rescue cells from β-defensin activity (Yan et al., 2013). Defensins can also act as chemoattractants for monocytes, dendritic cells, and selected lymphocytes (Faro-Trindade and Brown, 2009).

Another important antimicrobial peptide is LL-37, that kills *C. albicans* by fragmenting the cellular membrane of the fungus, leading to efflux of molecules like ATP and proteins (Den Hertog et al., 2005; Faro-Trindade and Brown, 2009). LL-37 can further act as a chemoattractant for neutrophils, monocytes and lymphocytes, induce histamine release from mast cells, alter the transcriptional response in macrophages and play a role in wound repair (Faro-Trindade and Brown, 2009). The peptide is produced by the proteolytic cleavage of cathelicidin (hCAP-18) (Den Hertog et al., 2005; Faro-Trindade and Brown, 2009). The hCAP-18 produced in neutrophils, and other cells including monocytes, natural killer cells, lymphocytes and a variety of epithelial cells, has antimicrobial activity itself (Faro-Trindade and Brown, 2009).

A family of cationic serine proteases called serprocidins also possess antimicrobial activity. Members of this family are stored within neutrophil granules and include protease-3, cathepsin G, and elastase. Those proteins are involved in many cellular processes including the cleavage of hCAP-18, cellular activation, as well as chemotaxis (Faro-Trindade and Brown, 2009).

Another example of an antimicrobial enzyme is lysozyme which targets the cell membrane of *C. albicans*. Lysozyme is expressed by a variety of phagocytes, including granulocytes, monocytes as well as macrophages and can be found at high levels in various tissues and secretions such as saliva. Its fungicidal activity is thought to occur through enzymatic hydrolysis of N-glycosidic bonds within the fungal cell wall and injury to the cell membrane (Faro-Trindade and Brown, 2009).

It is worth noting that the host makes use of *C. albicans*' polyamine influx transporters to get some antimicrobial peptides like histatin 5 into the pathogen. *C. albicans* in turn, uses multi-drug resistance pumps such as the fungal polyamine efflux transporter Flu1 to transport those antimicrobial peptides out again and thus reduce their toxicity (Li et al., 2013). *C. albicans* is also able to cleave histatin 5 with its protease Sap9 (Szafranski-Schneider et al., 2012).

Another mechanism by which *C. albicans* deals with antimicrobial peptides is the shedding of a large glycosylated fragment of Msb2. *C. albicans*' Msb2 stabilizes the fungal cell wall and inactivates histatin 5 and LL-37 (Szafranski-Schneider et al., 2012; Swidergall et al., 2013) as well as human α- and β-defensins (Swidergall et al., 2013).

2.2.2. Complement System

The complement system is highly efficient in recognizing and eliminating infectious pathogens while its activation is tightly regulated in time and space. For reviews about the interactions of *C. albicans* with the human complement system see Zipfel et al. (2011); Cheng et al. (2012); Zipfel et al. (2013); Luo et al. (2013).

For complement evasion *C. albicans* acquires several human complement regulators, e.g., C4BP (complement component 4b-binding protein), factor H, FHL-1 (four and a half LIM domains protein 1), plasminogen and vitronectin, to its cell surface to inhibit the actions of the complement system (Luo et al., 2013). Factor H is bound by four *C. albicans* proteins: phosphoglycerate mutase (Gpm1), Pra1, the high-affinity glucose transporter 1 (Hgt1p) and Gpd2 (Luo et al., 2013; Zipfel et al., 2013). *C. albicans* Pra1 and Hgt1p also bind C4BP. There are eleven *C. albicans* proteins that bind host plasminogen: Gpm1, enolase, Tsa1, Cta1 (catalase 1), Tdh3 (triose phosphate dehydrogenase 3), Tef1 (translation elongation factor 1-alpha), Pgk1 (phosphoglycerate kinase 1), Adh1 (alcohol dehydrogenase 1), Fba1 (fructose-bisphosphate aldolase), Pra1 and Gpd2 (Zipfel et al., 2013) and three *C. albicans* proteins which bind human FHL-1: Gpm1, Pra1 and Gpd2 (Luo et al., 2013). Human plasminogen bound on *C. albicans*' cells, can be activated to proteolytically active plasmin that cleaves host fibrinogen thereby contributing to the tissue invasion of *C. albicans* cells into epithelia cell layers (Zipfel et al., 2011). *C. albicans* further expresses $\alpha v\beta 3$ integrin-like protein that acquires host vitronectin to the fungal cell surface. This in turn inhibits the formation of the terminal complement complex (Luo et al., 2013). *C. albicans* can furthermore secrete aspartyl proteases Sap1, Sap2, and Sap3 that degrade the host complement proteins C3b, C4b and C5 (Gropp et al., 2009; Luo et al., 2013). The expression of endogenous complement inhibitors like secreted Pra1 which binds the central complement component C3 in solution is another mechanism by *C. albicans* to block C3 and complement activation (Zipfel et al., 2011). Secreted Pra1 also blocks the human integrin receptors CR3 and CR4, expressed by human leukocytes, granulocytes, macrophage and natural killer cells thereby inhibiting recognition, phagocytosis and cell-mediated killing (Luo et al., 2013).

Phagocytes respond to pathogens by recognizing opsonins and pathogen-associated molecular pattern (PAMPs) using surface expressed pattern recognition receptors (PRRs) (Jacobsen et al., 2012; Lopez, 2013; Luo et al., 2013). As the cell wall of *C. albicans* contains carbohydrates and cell wall proteins that are not present in the human body, it represents an ideal immunological target (Gow et al., 2012). Exhaustive reviews on *C. albicans*' cell wall architecture and its recognition have been published by Netea et al. (2008); Netea and Maródi (2010); Moyes and Naglik (2011); Gow et al. (2012). The most important *C. albicans* PAMPs are its cell wall carbohydrates: mannan (as mannosylated proteins), β-glucan, and chitin (Lopez, 2013). One mechanism used by *C. albicans* to evade the cellular

response by phagocytes is to shield these β-glucans with surface mannans upon hyphal growth to avoid phagocytosis (Luo et al., 2013). The receptor ligation of PRRs with PAMPs activates resident phagocytes and leads to synthesis and secretion of cytokines and lipid mediators. One evasion mechanism by *C. albicans* is the induction of an anti-inflammatory cytokine release by favoring toll-like receptor (TLR) 2 instead of TLR4 recognition (Zipfel et al., 2011). *C. albicans* is further able to inhibit the proinflammatory IL-17 production by altering the host tryptophan metabolism. This metabolism is regulated by two distinct enzymes: Indoleamine 2,3- dioxygenase (IDO) and tryptophan hydroxylase. By inhibiting IDO expression, *C. albicans* can shift the tryptophan metabolism, leading to fewer kynurenines and more 5-hydroxytryptophan metabolites. The increased 5-hydroxytryptophan levels subsequently inhibit the host IL-17 production (Cheng et al., 2010, 2012). A similar mechanism is used by cancer cells (Uyttenhove et al., 2003) and has been described by a mathematical model (Stavrum et al., 2013). For a detailed explanation of the recognition of *C. albicans* PAMPs and the *C. albicans* evasion strategies from epithelial cell defense see Netea et al. (2008); Netea and Maródi (2010); Moyes and Naglik (2011); Gow et al. (2012); Mech et al. (2014).

2.2.3. Phagocytes

While viral infections are primarily fought by T-cells in particular T-killer-cells, defense against fungi resembles bacteria defense in mobilizing neutrophils and macrophages.

Invading and disseminating *C. albicans* cells are faced with phagocytic cells (Kumar and Sharma, 2010; Zipfel et al., 2011; Jacobsen et al., 2012). Phagocytes, especially neutrophils and macrophages are of major importance for the host defense against mucosal and disseminated candidiasis (Cheng et al., 2012; Krysan et al., 2014; Quintin et al., 2014). These immune cells most effectively control and clear *C. albicans* infections by killing *C. albicans* cells intracellularly and extracellularly (Cheng et al., 2012). *C. albicans* on the other hand has evolved several mechanisms to control and evade the antimicrobial activity of local and newly attracted phagocytic cells by inhibiting recognition, trafficking, and effector release, thus overcoming several important stresses (Lopez, 2013; Luo et al., 2013).

Neutrophils are the prevalent immune cell type in anti-*Candida* immunity (Moyes and Naglik, 2011; Luo et al., 2013). During *C. albicans* infection, neutrophils migrate to sites of infection and release one or more chemotactic factors (Luo et al., 2013). After recognition of *C. albicans* cells through dectin-1 (recognizes β-1,3 glucan), dectin-2 (recognizes mannan), TLR2 (recognizes phospholipomannan), TLR4 (recognizes O-mannan), and mannose receptor (recognizes N-mannan) (Moyes and Naglik, 2011) neutrophils induce epithelial cell mediated protection against *C. albicans* infections and can directly kill *Candida* cells (Moyes and Naglik, 2011). The presence of neutrophils further inhibits *C. albicans* growth, including the yeast-to-hyphal transition (Jacobsen et al., 2012). These immune cells preferentially target *C. albicans* hyphae but kill yeast and hyphal forms of *C. albicans* at the same rate (Jacobsen et al., 2012; Tyc et al., 2014). Neutrophils rely on a range of antimicrobial effector mechanisms including oxidative burst, cytokine release, phagocytosis, neutrophil extracellular traps (NETs), release of granule enzymes as well as antimicrobial peptides to kill the fungus (Luo et al., 2013). Additionally they may differentiate into discrete subsets defined by distinct phenotypic and functional profiles (Scapini and Cassatella, 2014).

Another important immune cell type in anti-*Candida* immunity are macrophages (Jiménez-López and Lorenz, 2013; Krysan et al., 2014; Liu et al., 2014). These dynamic cells are distributed in various tissues and are part of the first line of host defense (Brunke and Hube, 2013; De Lima et al., 2014; Liu et al., 2014). Macrophages are of particular importance as they can both limit *C. albicans* burden early in infection and recruit and activate other immune effector cells (Krysan et al., 2014). Macrophages produce a variety of pro- and anti-inflammatory cytokines and chemokines in response to *C. albicans* (Jacobsen et al., 2012; Krysan et al., 2014). Particularly, *C. albicans* hyphae formation is a strong trigger for the production of IL-1β (Krysan et al., 2014) thereby helping to orchestrate the immune responses of the host (Jacobsen et al., 2012; Brunke and Hube, 2013). Cheng et al. (2011) showed that the development of hyphae during tissue invasion triggers the recognition by macrophages via the dectin-1/inflammasome pathway, leading to IL-1β production and thus T helper cell 17 (Th17 cell) activation. For a review on inflammasome activation see van de Veerdonk et al. (2015). Macrophages damage or directly kill *C. albicans* (Krysan et al., 2014) utilizing a combination of oxidative and nonoxidative microbicidal mechanisms including the production of antimicrobial peptides and degradative enzymes, the generation of ROS and nitric oxide synthase (iNOS), phagocytosis and macrophage extracellular traps (METs) (Liu et al., 2014). During phagocytosis macrophages readily ingest the round yeast form of *C. albicans* as well as relatively short filaments (Jacobsen et al., 2012; Brunke and Hube, 2013; Krysan et al., 2014). The fungus on the other side has developed several defense strategies to escape from macrophages with a significant cytotoxic effect on the immune cell, e.g., pyroptosis (Krysan et al., 2014).

Natural killer cells are innate lymphocytes with a potent cytotoxic activity. They usually are of major importance in viral infections and anti-tumor immunity (Voigt, 2013). The role of natural killer cells in host defense against *C. albicans* infection strongly differs depending on the state of host defense. While natural killer cells are an essential and non-redundant component of anti-*C. albicans* host defense in immunosuppressed hosts with defective T- and B-lymphocyte immunity they can contribute to hyperinflammation in immunocompetent hosts (Quintin et al., 2014). Natural killer cells modulate the immune responses by secreting cytokines which in turn recruit and activate other innate immune cells. Natural killer cells are also able to phagocytose *C. albicans* cells. However, in contrast to the professional phagocytic activity of neutrophils, this does not inhibit the further elongation of *C. albicans* filaments and leads to the destruction of the natural killer cell (Voigt, 2013). It was therefore proposed by Voigt (2013) that these immune cells contribute to the protective immunity against *C.*

albicans by recruiting other immune cells and enhancing proinflammatory activities without efficiently restricting the fungus.

Another important innate immune cell type for the *C. albicans* defense are dendritic cells. They are professional antigen-presenting cells which coordinate the immune response and link innate and adaptive immunity (Cheng et al., 2012; Ramirez-Ortiz and Means, 2012). Dendritic cells reside and patrol in the skin and mucosal surface and ingest *Candida* once tissues are invaded (d'Ostiani et al., 2000; Cheng et al., 2012). These immune cells use C-type lectin pattern recognition receptors like Dectin-1, Dectin-2 and DC-Sign to recognize the fungus. Dendritic cells phagocytose both yeast and hyphal *C. albicans* cells but kill yeast cells more efficiently (Jacobsen et al., 2012). After processing *C. albicans* they present *Candida*-specific antigens via major histocompatibility complex class II molecules (Cheng et al., 2012). Dendritic cells therefore have a bridging effect between the innate and adaptive antifungal responses. They are able to discriminate between yeast- and hyphal- forms of *C. albicans* (d'Ostiani et al., 2000; Cheng et al., 2012) and induce different T helper cell differentiation depending on the morphology of phagocytosed *C. albicans* cells (d'Ostiani et al., 2000; Cheng et al., 2012; Jacobsen et al., 2012). While yeast cells stimulate the priming of Th1 cells, the ingestion of hyphae inhibits IL-12 and Th1 differentiation, favoring Th2 cell differentiation (Cheng et al., 2012). The Th1 and Th17 cell responses are thought to be beneficial for the host (Jacobsen et al., 2012). The different responses of dendritic cells to yeast and hyphae morphologies may thus strongly influence the clinical course of infection (Hamad, 2012; Jacobsen et al., 2012).

2.2.4. Inside the Phagosome

Phagocytosis is depending on the glycosylation status of the *C. albicans* cell wall, the morphology of the fungus, the hyphal length, orientation and contact of the hyphae relative to the phagocyte as well as the immune cell types and their state of activation (Whittington et al., 2014). *C. albicans* in turn has developed mechanisms to resist phagocytic killing by escaping and even killing some phagocytic cell types (Faro-Trindade and Brown, 2009; Dementhon et al., 2012; Luo et al., 2013; Vylkova and Lorenz, 2014; Wellington et al., 2014). Several phagocytes can efficiently ingest *C. albicans* yeast cells and short hyphae (Jacobsen et al., 2012; Smith and May, 2013; Whittington et al., 2014). Accordingly, a natural evasion strategy is to form long hyphae because they can not be phagocytosed for simple geometrical reasons. This is analogous to needle-shaped microparticles which can not be engulfed either. Without intervention by the phagocytosed fungus, the phagosome matures via a series of fusion and fission events with the lysosome into the phagolysosome (Cheng et al., 2012; Brunke and Hube, 2013). However, early upon phagocytosis *C. albicans* is able to alter intracellular membrane trafficking within the phagosome by inhibiting phagosome maturation (Cheng et al., 2012; Dementhon et al., 2012; Vylkova and Lorenz, 2014). Inside the hostile environment of the phagolysosome *C. albicans* cells are killed and degraded by nutrient starvation, low pH levels, hydrolytic enzymes, antimicrobial peptides, ROS and reactive nitrogen species (NOS) (Faro-Trindade and Brown, 2009; Zipfel et al., 2011; Cheng et al., 2012; Luo et al., 2013; Mayer et al., 2013).

While neutrophils can block hyphal development (Faro-Trindade and Brown, 2009), *C. albicans* cells are able to generate hyphae within the phagolysosome of dendritic cells and macrophages allowing the fungus in some cases to kill and escape from those phagocytes (Faro-Trindade and Brown, 2009; Luo et al., 2013). Hyphal formation is depending on the pH-level (Faro-Trindade and Brown, 2009). While the acidic pH inside the phagosome should inhibit germination *C. albicans* is able to modulate the phagosomal milieu (Vylkova and Lorenz, 2014). *C. albicans* can rapidly alkalize the phagosomal environment via the arginine biosynthetic pathway (Lopez, 2013; Vylkova and Lorenz, 2014). Neutralization of the pH via the extrusion of ammonia presumably derived from the amino acid, results in the auto-induction of hyphal formation (Bain et al., 2012; Vylkova and Lorenz, 2014).

Some macrophages are able to withstand the stress of elongating *C. albicans* filaments without apparent loss of integrity (Krysan et al., 2014). In other macrophages, however, the *C. albicans* hyphal formation can provoke pyroptosis by activating the NLRP3 (NOD-like receptor family, pyrin domain containing 3) inflammasome and caspase-1. This proinflammatory, inflammasome-mediated programmed cell death pathway leads to the macrophage lysis and production of IL-1β and IL-18, allowing *C. albicans* to escape the hostile environment of the phagocyte. Early upon phagocytosis the majority of macrophage lysis is mediated by pyroptosis (Uwamahoro et al., 2014; Wellington et al., 2014). Later, a second macrophage killing phase, independent and distinct from pyroptosis, is initiated by *C. albicans* which depends on robust hyphal formation. As pyroptosis has a protective role in infections with bacterial pathogens by increasing inflammatory responses this might also be the case in *C. albicans* infections (Uwamahoro et al., 2014).

In a minority of cases phagocytosed *C. albicans* cells can escape from macrophages through non-lytic expulsion (Brunke and Hube, 2013; Lopez, 2013), also called exocytosis (Bain et al., 2012) or vomitosis (Whittington et al., 2014). This rare event is reported to occur at a low frequency but repeatedly in various experimental conditions (Bain et al., 2012). Although the underlying mechanisms are unknown (Lopez, 2013) it is observed that both, the macrophage as well as the *C. albicans* cell, remain intact and viable during phagocytosis and subsequent expulsion. This means the macrophage cell retains its phagocytic ability and is able to undergo mitosis as shown in Bain et al. (2012) while the *C. albicans* cell can perform hyphae elongation at normal rates. As non-lytic expulsion avoids lysis of the macrophage but also of the *C. albicans* cell this process may benefit both cell types (Bain et al., 2012). It is therefore not trivial to account the process as strategy to either macrophages or *C. albicans*.

2.2.5. Oxidative and Nitrosative Stress and its Detoxification

Phagocytes can produce oxidative and nitrosative stresses to kill *C. albicans* (Wilson et al., 2009; Cheng et al., 2012; Mayer et al., 2013) (see **Figure 1**). The respiratory burst as summarized by Faro-Trindade and Brown (2009) is mediated via the phagocyte NADPH oxidase (Phox). This membrane-associated protein complex generates superoxide through the transfer of electrons from NADPH to O_2. The generated superoxide (O_2^-) has little, if any, toxicity but can be converted to hydrogen peroxide (H_2O_2) and hydroxyl radicals (HO^-) with candidacidal activity. Myeloperoxidase (MPO), an enzyme located in granules of neutrophils and in lysosomes of monocytes (and even macrophages when they scavenge it through their mannose receptors), catalyzes the further conversion of hydrogen peroxide to hypochlorous acid (HClO). The hypochlorous acid in turn is an extremely toxic and effective candidacidal oxidant. The production of nitric oxide (NO) is induced by the inducible nitric oxide synthase (iNOS or NOS2) through the oxidative deamination of L-arginine. The NO itself has poor candidacidal activity but can further react with the superoxide (O_2^-), generated by the respiratory burst (Vazquez-Torres et al., 1996; Faro-Trindade and Brown, 2009). The so produced peroxynitrite ($ONOO^-$), an unstable structural isomer of nitrate (NO_3^-), is very effective at killing *C. albicans* (Vazquez-Torres et al., 1996; Faro-Trindade and Brown, 2009; Cheng et al., 2012). As the production of ROS and nitrosative stress are major antifungal mechanisms in phagocytes, *C. albicans* possesses several defense strategies to counteract the oxidative and nitrosative stresses (Cheng et al., 2012; Mayer et al., 2013).

The response of *C. albicans* to ROS is regulated by Cap1 (adenylate cyclase-associated protein 1) and the MAP (mitogen-activated protein) kinase Hog1. Both proteins regulate the catalase expression in *C. albicans* (Lopez, 2013). The fungus can produce antioxidant enzymes like the catalase Cta1 and intracellular as well as extracellular superoxide dismutases (Sods) to counteract the respiratory burst (Faro-Trindade and Brown, 2009; Frohner et al., 2009; Cheng et al., 2012; Mayer et al., 2013; Miramón et al., 2013). Of the intracellular Sods, Sod1 is required for interaction with macrophages and Sod2 is necessary to resist neutrophil attack (Miramón et al., 2013). Next to the catalase Cta1, the superoxide dismutases Sod5, Sod4 (Frohner et al., 2009) and Sod6 detoxify extracellular ROS produced by macrophages (Lopez, 2013). The expression of Sods is depending on the fungal morphology. While Sod4 is expressed by *C. albicans* yeast cells, the hyphal forms express Sod5 (Miramón et al., 2013). Neutrophils also induce the expression of Sod5 even though they inhibit the yeast-to-hyphal formation in *C. albicans* (Frohner et al., 2009; Miramón et al., 2013). Furthermore, in response to incubation with neutrophils Sods, catalase, glutathione peroxidase, glutathione reductase, glutathione S-transferase and thioredoxin are strongly induced in *C. albicans* cells (Wilson et al., 2009). The superoxide detoxification generates H_2O_2 which is still highly toxic but subsequently eliminated by Cat1. The glutathione peroxidases (Gpxs) also detoxify H_2O_2 via oxidation of the thiolgroups in

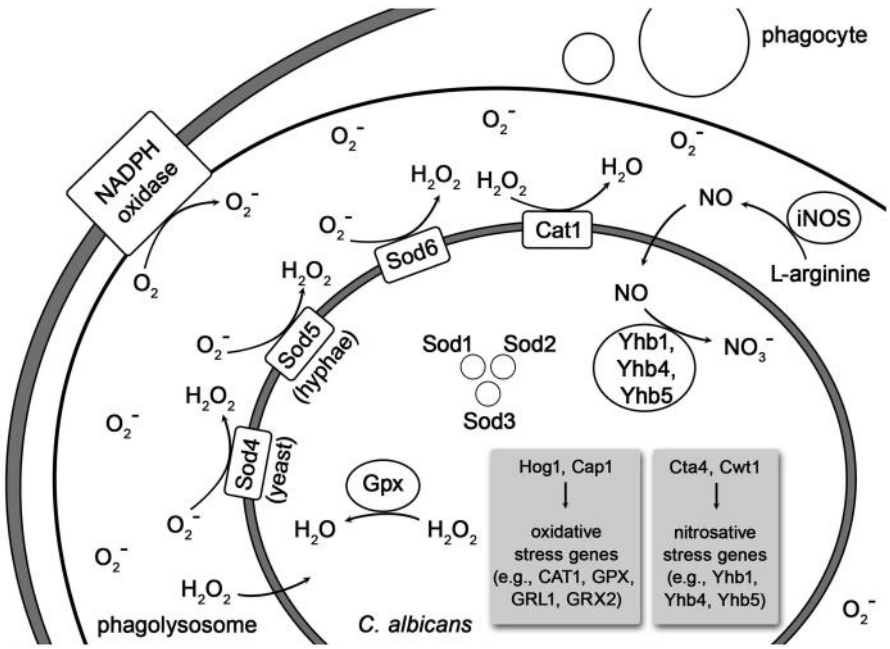


FIGURE 1 | Depiction of oxidative and nitrosative stress imposed on phagocytosed *C. albicans* and its detoxification by the fungus. The Curved lines indicate cell membranes of the phagocyte, the phagolysosome and, most inside, the *C. albicans* cell. Abbreviations: iNOS, inducible nitric oxide synthase; Cta1, catalase 1; Sod1-6, superoxide dismutases 1–6; Gpxs, glutathione peroxidases; GRX2 and GRL1 encode glutathione reductases; Yhb1, Yhb4, and Yhb5, flavohemoglobin 1, 4, and 5; Hog1, mitogen-activated protein kinase; Cap1, adenylate cyclase-associated protein; Cta4, transcription factor; Cwt1, cell wall transcription factor.

two glutathione molecules, which are subsequently reduced by glutathione reductases (Grxs), encoded by GRX2 and GRL1 (Miramón et al., 2013). In addition *C. albicans* up-regulates DNA damage repair systems and heat shock proteins to counteract oxidative damage to nucleic acids and proteins (Faro-Trindade and Brown, 2009). The exposure to moderate concentrations of ROS induces the entire arginine biosynthetic pathway but no other amino acid synthetic genes in phagocytosed *C. albicans* cells (Lopez, 2013).

The nitrosative stress response of *C. albicans* is mediated by the three intracellular flavohemoglobin enzymes Yhb1, Yhb4, and Yhb5 (Lopez, 2013; Mayer et al., 2013) which convert NO to less toxic NO_3^- molecules (Luo et al., 2013). The nitrosative stress response is regulated by the two transcription factors Cta4 and Cwt1. While Cta4 positively regulates the transcriptional nitrosative stress response, Cwt1 negatively regulates it (Miramón et al., 2013).

2.2.6. Extracellular Traps

The production of extracellular traps (ETs) is a phagocytosis independent antimicrobial mechanism observed in many effector cells including neutrophils and macrophages (Pruchniak et al., 2012; Branzk and Papayannopoulos, 2013; Hahn et al., 2013; Liu et al., 2014). These fiber-like extracellular structures are induced by many different microbes including *C. albicans* (Faro-Trindade and Brown, 2009), chemicals and cytokines (Liu et al., 2014). As being significantly associated with the microbial surface ETs are thought to act as a physical barrier that prevents invading pathogens from further progressing. The formation of ETs was therefore proposed as supplementary strategy by the host defense when phagocytosis failed to eliminate the invading pathogen (Liu et al., 2014).

Neutrophil extracellular traps (NETs) occur as specialized form of neutrophils cell death and consist of DNA scaffolds with antimicrobial proteins like histones and granule proteins including myeloperoxidase, elastase, cathelicidins, cathepsin G, calprotectins and gelatinase B (Faro-Trindade and Brown, 2009; Urban et al., 2009; Moyes and Naglik, 2011; Liu et al., 2014). Releasing these effector molecules into the extracellular space allows neutrophils to efficiently trap and kill the yeast and hyphal forms of *C. albicans* (Faro-Trindade and Brown, 2009; Liu et al., 2014). While this is beneficial for the host defense, NETs also participate in propagating some autoimmune diseases such as systemic lupus erythematosus and small vessel vasculitis (Liu et al., 2014).

C. albicans cells also induce the formation of METs like structures (METs-LS). These METs-LS can be released by dying as well as viable macrophages and thus show more than one type of composition. While some METs-LS consist of a DNA backbone and microbicidal proteins including histone, myeloperoxidase and lysozyme, other METs-LS did not contain histone. As histones are associated with nuclear DNA it was proposed that the DNA backbone in those METs-LS without histone originates from mitochondrial DNA. In contrast to NETs, METs-LS are not capable to efficiently kill *C. albicans* cells. Instead METs-LS rather contain the invading pathogen at the infection site, thereby preventing the systemic diffusion of *C. albicans* and providing time to recruit other effector cells like neutrophils (Liu et al., 2014).

While the formation of ETs usually depends on the generation of ROS via the activation of the NADPH oxidase, *C. albicans* induces NETs and METs-LS in an ROS independent manner (Liu et al., 2014).

3. Computational Systems Biology Approaches

In many fields of biology, Computational Systems Biology approaches have turned out to be very useful (Heinrich and Schuster, 1996; Klipp et al., 2011). Various Systems Biology methods for understanding and predicting fungal virulence have been reviewed by Tierney et al. (2014). For other organisms, it has been shown that network analyses are useful to describe and understand the manifold interactions between a pathogen and its host (Naseem et al., 2012).

The basis for many Systems Biology approaches is provided by high throughput data. There are several studies regarding the omics of *Candida*. For instance, eight *Candida* genomes were compared by Butler et al. (2009). They found large families and genome expansions regarding the cell wall, secreted proteins and transporters, in particular in pathogenic species. These adaptations seemed thus to be associated with virulence.

Comprehensive transcriptome data were collected by Bruno et al. (2010). Measuring gene expression they identified 602 novel transcriptionally active regions. Conditions included hyphae-induction, tissue culture, high and low oxidative stress, nitrosative stress as well as cell wall damage-inducing conditions.

Regarding the omics of infection, there are in principle also dual sequencing approaches feasible but this is not really explored yet. Instead, Liu et al. (2015) investigated the host response to *C. albicans* infection in various niches and derived exciting results. Network analysis, siRNA knock down and RNAseq data identified new host signaling pathways under infection such as platelet-derived growth factor BB (PDGF BB) and neural precursor-cell-expressed developmentally down-regulated protein 9 (NEDD9). Both proteins regulate the uptake of *C. albicans* by host cells.

Regarding metabolite data, there is a lipidomics study by Singh et al. (2013) studying changes in *C. albicans* caused by fluconazole, a drug against candidiasis. Under this treatment, *C. albicans* shows an increased sterol content and depleted sphingolipid levels in case of azole resistance.

As evidenced by omics data many mechanisms and phenomena in biology (e.g., entangled positive and negative feedback loops) are so complex that they cannot be understood by intuition. This is one reason for the ever increasing importance of computer simulations. A first and important step is to explain known phenomena on theoretical grounds, thus helping us to understand them. The usefulness of this aspect of Computational Systems Biology should not be underestimated. A famous early example is the Michaelis-Menten kinetics. This formal approach helps us to understand the role of the association and dissociation of the enzyme-substrate complex.

It has a predictive aspect because it allows one to calculate the reaction velocity even for substrate concentrations for which no measurement has been performed for a specific enzyme so far.

The most ambitious goal is to predict hitherto unknown properties, interactions and behaviors. Several studies show that Computational Systems Biology approaches can generate clear testable predictions that could later be confirmed in experiments (Schuster et al., 2006) or highlight new working hypotheses for *in vitro* experiments (Siegismund et al., 2014a,b). The latter study investigates the early colonization of bacteria on different biomaterials typically used for implants. Automated images analysis of CLSM images and point-pattern analysis were applied to show material-induced switches from bacterial adhesion to colony growth on biomaterials. By two- or three-dimensional modeling, e.g., using cellular automata or agent-based models, the adhesion of pathogens and/or epithelial cells of the host on implant surfaces can be simulated. For the case of bacteria, see Siegismund et al. (2014b). This helps us to understand the onset of disease in the case where the pathogens win this "race for the surface" (Subbiahdoss et al., 2009) and the avoidance of disease in the case where the host cells win. This will also help to devise novel therapeutic strategies, as an appropriate surface structure of the implants can diminish adhesion by pathogens. A model for the thermal adaptation of *Candida* based on a differential equation system was proposed by Leach et al. (2012). That model appropriately describes the defense-evasion pair "fever—heat shock response." Moreover, other modeling techniques have been used to study *Candida* infections, such as Bayesian modeling (Shankar et al., 2015) and dynamic interactive infectious networks (Chen and Wu, 2014).

Several defense and evasion mechanism have been described by mathematical modeling. For example, the action of degradative enzymes can be simulated by kinetic models of metabolic networks (Heinrich and Schuster, 1996). Kinetic models of tryptophan metabolism (Stavrum et al., 2013) and of multi-drug resistance pumps have been published (Westerhoff et al., 2000). A large body of literature on the modeling of biofilm formation is available (Audretsch et al., 2013), though mostly on bacterial rather than fungal biofilms, for a review see Horn and Lackner (2014). Moreover, a gene regulatory network was inferred (Tierney et al., 2012). All of those modeling techniques could in principle be applied to investigate *C. albicans*' interactions with the host.

In the present chapter, we outline the modeling methods based on game theory and agent-based models in more detail.

3.1. Game theory

Metaphorically, the struggle between pathogens and the human immune system can be considered as a game in which each player attempts to win (Renaud and De Meeus, 1991; Hummert et al., 2014). This metaphor is quite useful because it allows one to understand that struggle as an extended optimization process. The extension is that the two counterparts (players) cannot always reach the optimal state because they may hinder each other in reaching it. Thus, suboptimal states can result (Hofbauer and Sigmund, 1998). A considerable number of game-theoretical models of bacterial and viral infections have been proposed, for a review see Hummert et al. (2014), while fungal infections are the subject of such studies to a lesser extent so far. To our knowledge, Hummert et al. (2010) were the first to present a game-theoretical model of the interaction of *C. albicans* with the human immune system, in particular, with human macrophages. The simplifying assumption used was that *C. albicans* has two strategies when engulfed by macrophages: avoiding lysis transiently (silencing) or undergoing a morphological switch to form hyphae and escaping (piercing). The latter situation corresponds to the defense-evasion pair "phagocytosis—pyroptosis." In the approach by Hummert and coworkers, different *Candida* cells are considered as players while the macrophage was considered as a constant environment. Thus, a symmetric game results and the fitness matrix can be written as follows:

	p	s
p	$l-c$	$l-c$
s	$\left(1-e^{-\lambda}\right) l$	0

where s and p stand for the silencing and piercing strategies, respectively, l stands for the benefit of surviving, c for having payed the costs for piercing and λ denotes the average number of engulfed cells. A Poisson distribution for the number of ingested *C. albicans* yeast cells was assumed.

Every entry in the fitness matrix gives the payoff of an individual playing a row strategy against a pure population playing the column strategy. Under certain parameter conditions, a pure piercing population can exist. For other parameter values, a mixed evolutionary stable strategy (ESS) results, which corresponds to coexistence of silencing and piercing cells. The silencing cells then benefit from the efforts made by the piercing cells. In game-theoretical terms, this is a hawk-dove game (Hofbauer and Sigmund, 1998; Stark, 2010). Both of the above-mentioned outcomes are in good agreement with experimental observations, because two different karyotypes had been found (Tavanti et al., 2006).

A related model was established to describe the switch from yeast to hyphae upon invasion of human tissues (rather than inside macrophages) (Tyc et al., 2014). These authors extended the model by differential equations, allowing them to describe the dynamic behavior. Two situations are compared: cooperation between yeast and hyphae forms, meaning that the yeast form will also benefit when some cells switch to become hyphae, and competition, in which coexistence of yeast and hyphal cells pays off only to the hyphae. The model predicts that cooperation among fungal cells occurs in mild infections and an enhanced tendency to invade the host is associated with the competitive behavior (Tyc et al., 2014).

Recent progress investigated the iterative Prisoners' Dilemma. Interestingly, there is an incentive for cooperativity under these circumstances. It is worthwhile to use these mathematical

insights in the context of recurrent *Candida* infections (an often happening medical condition). As known for different bacterial strains such as *Pseudomonas* in Mucoviscidosis there should be some signs for selection for mitigated strains in such repeated *Candida* infections. Furthermore, dictator strategies force a certain win or loss on the opponent, no matter which strategy is chosen (Axelrod and Hamilton, 1981). Such a way of action should be the typical strategy of the immune system in the healthy person but has not yet been extensively investigated, in particular as such healthy persons rarely undergo clinical investigation.

In contrast to the game between different *Candida* cells, the "game" between the immune system and pathogens is an asymmetric game. A pioneering paper on that type of description was published by Renaud and De Meeus (1991) for the general case of any pathogen. The two players can choose between an aggressive strategy (called the "killer" strategy) that seeks to eliminate the adversary and a less aggressive strategy ("diplomat").

Renaud and De Meeus (1991) wrote down a rather general payoff matrix involving several parameters. To illustrate the idea, we here give a more specific matrix. However, the concrete numbers do not matter as long as they fulfill certain order relations. The entries in the matrix can be explained as follows. If both sides adopt the "killer" strategy, they win with a certain probability and have to afford the costs for that aggressive behavior. This is here quantified by 1 for either side. If both adopt the "diplomat" strategy, they can coexist and need not afford the costs for aggression. Thus, they can gain, say, 5 points each. If the host and parasite play "killer" and "diplomat," respectively, the latter will be eliminated (payoff of 0). The host survives but has to afford some costs, so that its payoff is between 1 and 5 (here assumed to be 3). In the converse situation, the host will die (or at least become very sick), here quantified by 0. The parasite has a benefit b and some cost c.

Parasite: Host:	Killer	Diplomat
Killer	1, 1	3, 0
Diplomat	0, $b-c$	5, 5

The type of game depends on the difference $b - c$. If it is less than 5, there are two stable Nash equilibria on the main diagonal: "killer, killer" and "diplomat, diplomat." Loosely speaking, we can denote them by "war" and "peace." The game is then related to the coordination game, in which the two players have to coordinate with each other to select among two symmetric Nash equilibria (Stark, 2010). Although the peaceful situation is better for both of them, they can get stuck in the war because neither side can leave it unilaterally without decreasing its payoff even more. It is worth noting that the non-lytic expulsion of *C. albicans* cells from macrophages mentioned in Section 2.2.4 can be considered as a peaceful situation as well.

If $b - c > 5$ (that is, high benefit or low cost of "killer" strategy for the parasite), peace is no longer stable because there is an incentive for the parasite to switch to the "killer" strategy. The state "diplomat, killer" is not, however, stable either because the host will then switch to "killer" as well. Thus, only war is stable. This change of Nash equilibria is a suitable model for the change from immunocompetent to susceptible hosts (e.g., after antibiotics treatment due to change in bacterial flora). The cost c for the parasite to invade the host then decreases, so that $b - c$ can exceed the critical threshold.

The situation where only war is stable in the killer-diplomat game is quite paradoxical because both sides would be better off if they were in peace with each other. This is reminiscent of the famous Prisoners' Dilemma, which is a symmetric game (Stark, 2010). In fact, the cause for instability is similar in both games: it is the temptation (incentive) to leave the mutually beneficial state. The difference between the two games is that, in the Prisoners' Dilemma, both players are tempted in this way while in the killer-diplomat game, there is a temptation for the parasite only.

3.2. Agent-based modeling

Agent-based models (ABMs) have become a powerful tool for tackling complex systems, where the individuality, temporal state and spatial distribution of its players may be of importance. They are typically characterized by numerous interacting entities, often called agents or individuals (depending on the discipline so that the term individual-based model (IBM) is used as well). They pursue certain objectives (e.g., increasing fitness, yield, status) by following, more or less, simple structured rules. These agents can be mobile or stationary units within a continuous or discrete environment defined by three, two, one or even no spatial dimension. *In silico* environments without any dimension simply imply that the modeled system behavior is presumably independent of any spatial scale. Including more dimensions assumes that this may be of importance for the behavior of the system: A model investigating the hunting strategies of a terrestrial predator may be sufficiently described by a two-dimensional environment. Whereas a third dimension has to be considered simulating the movement of immune-cells through different tissues or in the blood.

The philosophy of ABMs is to slice problems on the macro-level down to simple interaction- and reaction-rules of players on a micro-level. For example, patterns occurring on the population level are transferred to properties and the behavior of single individuals. Diseases of an individual can be explained by the malfunction or disorder of organs and tissues. Often the macro-level behavior of a system cannot be foreseen by only summing up the rules of players. Instead, patterns may arise from the complex interdigitation of state-dependent behavior of its entities, an effect called emergence.

Resolving a macro-level pattern (emergence of a certain behavior) to a lower complexity level comes at the price of a detailed knowledge of the individuals properties and behavioral strategies, which have to be precisely formulated. Especially models representing a biological system frequently deal with several involved types of agents (e.g., food-webs, stability of

ecosystems) and numerous interactions often require a bottom-up modeling approach with a deep knowledge of individual properties. Thus, ABMs are typically hungry for data (e.g., thresholds for reaction to signals, kinetic parameters) and computationally expensive due to, e.g., a frequent use of random number generators to induce local and individual stochasticity. Beside classical experimental approaches (e.g., for determining growth-rates), image- and video-derived data offer a valuable complementary solution to fill this gap of knowledge (Mech et al., 2011, 2014).

Deviating from an equation-based modeling approach typical ABMs show a considerable set of non-redundant rules. This often poses difficulties to communicate ABMs. Grimm and colleagues addressed this obstructive problem by proposing a standardized- and later updated protocol to formalize the descriptions of ABMs (Grimm et al., 2006, 2010).

Tokarski et al. (2012) investigated several hunting strategies of alveolar macrophages for fungal spores of *Aspergillus fumigatus*. The clearing efficiency of the immune system represented the emergent property of this system. Different scenarios of interactions between both players were tested, e.g., random walk of macrophages; detection and guidance of macrophages along local gradients of degradation products, indicating sporulation of spores and positive feedback activation of macrophages which already detected fungal spores. This approach exemplarily shows that biological systems above a certain degree of complexity, would be hard, if not impossible, to handle with an equation-based model. System properties may only arise at such a high complexity, e.g., by the local, state-dependent interactions of several agents (see **Figure 2**).

Due to its strengths in representing complex systems, ABMs show a broad field of scientific applications incorporating ecological (DeAngelis and Grimm, 2014) and microbial questions (Kreft et al., 2013).

ABM approaches helped to understand the epidemic spread of diseases, e.g., influenza (Milne et al., 2008; Laskowski et al., 2011) and Ebola (Merler et al., 2015) or, the microbial resistance to antibiotics in *Staphylococcus aureus* (Macal et al., 2014). But were also utilized to test the efficiency of counter-strategies for disease control (Borkowski et al., 2009; Tian et al., 2013; Havas et al., 2014).

Ideally ABMs can predict biological mechanisms or strategies which were unknown from wet-lab experiments, showing that both approaches are not competitive but complementary. Pollmächer and Figge (2014) investigated migration modes of macrophages and predicted that their efficiency of finding conidia could only be explained by the release of chemotactic signals from epithelial cells associated with *Aspergillus fumigatus*.

The human immune defense system consists of a dense mesh of state-dependent interactions between numerous types of players (e.g., pathogens, neutrophils or dendritic cells). Cascades of signal molecules are steering the induction and inhibition of cell responses and locally trigger the mobilization of different defense levels (e.g., passive, innate and acquired).

Folcik et al. (2007) examined the complex interplay between the innate and adaptive parts of the immune system. The focus was on the qualitative response to a viral infection. That work showed that all parts of the immune system are non-redundant and deficiency in any components increased the probability of failure to clear the simulated viral infection.

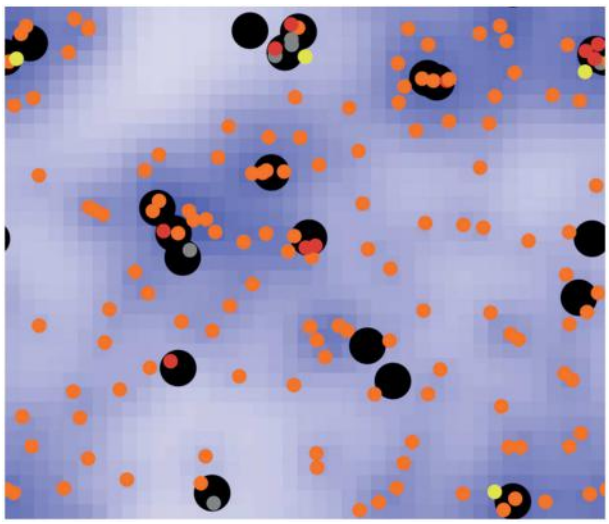


FIGURE 2 | Screenshot of a typical ABM simulation (taken from Tokarski et al., 2012), displaying the hunt of free-moving neutrophil agents (black circles) for immobile spore agents of *Aspergillus fumigatus* in human lung tissue. Spores can be free (orange), temporarily dragged (yellow) or caught (red) by a neutrophil agent. Dragged spores may be released with a certain probability or caught and phagocytosed (gray). Neutrophil agents are able to detect chemokines (blue), released by spores during sporulation, and may adjust their movement accordingly.

Baldazzi et al. (2006) investigated anti-HIV therapy with a immune-system model of multiple immune-cell types. ABMs typically show a high degree of specificity, thus representing one specific issue of a complex system in detail. Thus, ABMs are often less general and hard to transfer to similar questions. Examples addressing specific pathogens are: *Clostridium* (Peer and An, 2014), *Pseudomonas aeruginosa* (Seal et al., 2011), *Leishmania* (Dancik et al., 2010), *Helicobacter pylori* (Carbo et al., 2013), and *Aspergillus fumigatus* (Tokarski et al., 2012; Pollmächer and Figge, 2014). Tyc and Klipp (2011) suggested how to combine the complex behaviors of both, the host and the pathogen. Extensive reviews regarding ABMs of the immune system can be found in Chavali et al. (2008); Bauer et al. (2009); Li et al. (2008); Forrest and Beauchemin (2007).

4. Discussion

In this review, we have given an overview of the immune defense mechanisms of the human host against *C. albicans* and the evasion mechanisms of the fungus to escape, circumvent or counteract the immune response. Both the terms defense and evasion are here used in a wide sense and may include attack mechanisms. While earlier reviews have given an overview of experimental observations on *C. albicans* defense and evasion strategies, we present here an integrative synthesis of experimental observation and theoretical modeling of infection

strategies of *C. albicans*. Our review extends previous efforts on this topic (Zipfel et al., 2013).

On the basis of the list of mechanisms and strategies, given in **Tables 1**, **2**, it is of interest to search for even higher levels of interaction, that is, whether there are cascades including counter-counter defenses. For example *Streptomyces clavuligerus* produces both penicillin and clavulanic acid, a β-lactamases inhibitor (Reading and Cole, 1977; Knowles, 1985). Clavulanic acid and other β-lactamase inhibitors like sulbactam and tazobactam, limit the destructive action of β-lactamases from bacteria against β-lactam compounds such as penicillins and cephalosporins (Williams, 1997). Thus, there are three levels in the case of *Streptomyces clavuligerus*: penicillin as a defense chemical, β-lactamases as an evasion (counter-defense) mechanism by bacteria and clavulanic acid as a counter-counter defense. To our knowledge, no counter-counter defense is known in the case of *C. albicans* so far. However, Qiao et al. (2013) showed that other eukaryotic pathogens, i.e., oomycetes, are able to suppress RNA silencing, for a review see Pumplin and Voinnet (2013). Examples for counter-counter defense strategies are also known from plant-virus interactions, i.e., by antagonizing the virus-induced downregulation of RNA silencing by the plant (Sansregret et al., 2013). An intriguing question in the microbiology of pathogens is: How deep such an arms race of a host-pathogen interaction may evolve? Or, in other words: Are organisms rather selected for a counter-counter defense or an evolutionary novel mechanism of direct defense. The efficiency of multiple and complex layers of defense and counter-defense can be described mathematically by methods from Operations Research (Abt, 1987).

As the examples given above like macrophage phagocytosis and pyroptosis show, it can be hard to predict the outcome of the struggle between the human immune system and *C. albicans*. A special focus of our review therefore lies on the discussion of various Systems Biology approaches. Those are undoubtedly a promising tool to represent complex host-pathogen interactions and allow for the emergence of observed *in vivo* outcomes and for extensive testing scenarios (e.g., medication, drug testing, cross-effects). For example the acquisition of human complement regulators to the cell surface can be considered as a molecular mimicry. Mathematical models of mimicry in higher organisms can be adapted to describe this phenomenon. Systems Biology approaches are instrumental for questions which are hard to conduct solely in laboratory experiments. Nevertheless, theoretical approaches have to be substantially supported and completed by *in vivo* and *in vitro* approaches.

In **Tables 1**, **2**, both specialist and generalist effector mechanisms can be seen. One example for a generalist effector is the *C. albicans* protein Pra1 as it exerts several effects. In the terminology of networks analysis, Pra1 is a hub. Accordingly, it is of interest in future studies to analyze how complex and entangled the network of interactions is, whether it is scale-free or has small-world properties etc. (Yook et al., 2004). These properties are relevant in view of robustness against errors and mutations (Albert et al., 2000).

Another interesting question is how the host protects itself from its own "attack" mechanisms such as oxidative and nitrosative stress. Obviously, the levels of these substances should not exceed upper limits. This, in turn, might give a chance for *C. albicans* in its evasion strategies. Moreover, an optimal trade-off between immunity and autoimmunity as in the case of NETs must be found which also implies upper limits on the degree of defense. In this context, the camouflage by *C. albicans* using factor H is worth mentioning.

On or within the human host *C. albicans* not only interacts with the host but also with all the probiotic microorganisms of the host's microflora. It is therefore worthwhile to further look into these interactions, e.g., *C. albicans*' theft of iron from siderophores produced by other microorganisms via its own siderophore uptake system. Kleptoparasitism can be investigated using game theoretic models, considering individuals as well as groups (Broom and Rychtář, 2011). To our knowledge this has not been done for *C. albicans* so far.

All these topics are prone to be analyzed by mathematical modeling and computer simulations. Some of the computational methods such as agent-based models allow one to describe both temporal and spatial aspects. Ideally, *in-silico* modeling makes it possible to reduce the number of experiments with animals and ethically questionable or prohibited experiments with humans. This helps to gain further insight and make medically important predictions, for instance regarding onset of fungal sepsis and novel intervention strategies in the immunocompromised patient.

Author Contributions

Conception and design of the investigation and work: all. Drafting the manuscript: SD, SG, SS. Revising it critically for important intellectual content and final approval of the version to be published: all. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: all.

Funding

Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 "Pathogenic fungi and their human host: Networks of interaction" subproject B1 (SD, TD, and SS), subproject C4 (CS), and subproject C6 (PFZ).

Acknowledgments

The authors thank Maria Praußе for designing **Figure 1** and her help with the literature search as well as Hortense Slevogt, Stefan Lorkowski, and Karsten Gneist for stimulating discussions.

References

Abt, C. C. (1987). *Serious Games*. New York, NY: University Press of America.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382. doi: 10.1038/35019019

Audretsch, C., Lopez, D., Srivastava, M., Wolz, C., and Dandekar, T. (2013). A semi-quantitative model of quorum-sensing in *Staphylococcus aureus*, approved by microarray meta-analyses and tested by mutation studies. *Mol. Biosyst.* 9, 2665–2680. doi: 10.1039/c3mb70117d

Axelrod, R., and Hamilton, W. D. (1981). The evolution of cooperation. *Science* 211, 1390–1396. doi: 10.1126/science.7466396

Baillie, G. S., and Douglas, L. J. (1998). Effect of growth rate on resistance of *Candida albicans* biofilms to antifungal agents. *Antimicrobial Agents Chemother.* 42, 1900–1905.

Baillie, G. S., and Douglas, L. J. (1999). Role of dimorphism in the development of *Candida albicans* biofilms. *J. Med. Microbiol.* 48, 671–679. doi: 10.1099/00222615-48-7-671

Bain, J. M., Lewis, L. E., Okai, B., Quinn, J., Gow, N. A., and Erwig, L.-P. (2012). Non-lytic expulsion / exocytosis of *Candida albicans* from macrophages. *Fungal Genet. Biol.* 49, 677–678. doi: 10.1016/j.fgb.2012.01.008

Baldazzi, V., Castiglione, F., and Bernaschi, M. (2006). An enhanced agent based model of the immune system response. *Cell. Immunol.* 244, 77–79. doi: 10.1016/j.cellimm.2006.12.006

Bauer, A. L., Beauchemin, C. A. A., and Perelson, A. S. (2009). Agent-based modeling of host-pathogen systems: the successes and challenges. *Inf. Sci.* 179, 1379–1389. doi: 10.1016/j.ins.2008.11.012

Borkowski, M., Podaima, B. W., and McLeod, R. D. (2009). Epidemic modeling with discrete-space scheduled walkers: extensions and research opportunities. *BMC Public Health* 9(Suppl. 1):S14. doi: 10.1186/1471-2458-9-S1-S14

Branzk, N., and Papayannopoulos, V. (2013). Molecular mechanisms regulating NETosis in infection and disease. *Semin. Immunopathol.* 35, 513–530. doi: 10.1007/s00281-013-0384-6

Broom, M., and Rychtář, J. (2011). Kleptoparasitic melees-modelling food stealing featuring contests with multiple individuals. *Bull. Math. Biol.* 73, 683–699. doi: 10.1007/s11538-010-9546-z

Brunke, S., and Hube, B. (2013). Two unlike cousins: *Candida albicans* and *C. glabrata* infection strategies. *Cell. Microbiol.* 15, 701–708. doi: 10.1111/cmi.12091

Bruno, V. M., Wang, Z., Marjani, S. L., Euskirchen, G. M., Martin, J., Sherlock, G., et al. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res.* 20, 1451–1458. doi: 10.1101/gr.109553.110

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A., Sakthikumar, S., Munro, C. A., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459, 657–662. doi: 10.1038/nature 08064

Carbo, A., Bassaganya-Riera, J., Pedragosa, M., Viladomiu, M., Marathe, M., Eubank, S., et al. (2013). Predictive computational modeling of the mucosal immune responses during *Helicobacter pylori* infection. *PLoS ONE* 8:e73365. doi: 10.1371/journal.pone.0073365

Chandra, J., Kuhn, D. M., Mukherjee, P. K., Hoyer, L. L., McCormick, T., and Ghannoum, M. A. (2001). Biofilm formation by the fungal pathogen *Candida albicans*: development, architecture, and drug resistance. *J. Bacteriol.* 183, 5385–5394. doi: 10.1128/JB.183.18.5385-5394.2001

Chavali, A. K., Gianchandani, E. P., Tung, K. S., Lawrence, M. B., Peirce, S. M., and Papin, J. A. (2008). Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. *Trends Immunol.* 29, 589–599. doi: 10.1016/j.it.2008.08.006

Chen, B.-S., and Wu, C.-C. (2014). "A systems biology approach to study systemic inflammation," in *Immunoinformatics (Methods in Molecular Biology)*, eds R. K. De and N. Tomar (New York, NY: Springer), 403–416.

Cheng, S.-C., Joosten, L. A., Kullberg, B.-J., and Netea, M. G. (2012). Interplay between *Candida albicans* and the mammalian innate host defense. *Infect. Immun.* 80, 1304–1313. doi: 10.1128/IAI.06146-11

Cheng, S.-C., Van de Veerdonk, F., Smeekens, S., Joosten, L. A., Van der Meer, J. W., Kullberg, B.-J., et al. (2010). *Candida albicans* dampens host defense by downregulating IL-17 production. *J. Immunol.* 185, 2450–2457. doi: 10.4049/jimmunol.1000756

Cheng, S.-C., van de Veerdonk, F. L., Lenardon, M., Stoffels, M., Plantinga, T., Smeekens, S., et al. (2011). The dectin-1/inflammasome pathway is responsible for the induction of protective T-helper 17 responses that discriminate between yeasts and hyphae of *Candida albicans*. *J. Leukoc. Biol.* 90, 357–366. doi: 10.1189/jlb.1210702

Chin, V. K., Foong, K. J., Maha, A., Rusliza, B., Norhafizah, M., and Chong, P. P. (2014). Early expression of local cytokines during systemic *Candida albicans* infection in a murine intravenous challenge model. *Biomed. Rep.* 2, 869–874. doi: 10.3892/br.2014.365

Collette, J. R., and Lorenz, M. C. (2011). Mechanisms of immune evasion in fungal pathogens. *Curr. Opin. Microbiol.* 14, 668–675. doi: 10.1016/j.mib.2011.09.007

Curtis, M. M., and Way, S. S. (2009). Interleukin-17 in host defence against bacterial, mycobacterial and fungal pathogens. *Immunology* 126, 177–185. doi: 10.1111/j.1365-2567.2008.03017.x

Dancik, G. M., Jones, D. E., and Dorman, K. S. (2010). Parameter estimation and sensitivity analysis in an agent-based model of *Leishmania major* infection. *J. Theor. Biol.* 262, 398–412. doi: 10.1016/j.jtbi.2009.10.007

De Figueiredo, L. F., Schuster, S., Kaleta, C., and Fell, D. A. (2008). Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics* 24, 2615–2621. doi: 10.1093/bioinformatics/btn500

De Lima, T. M., Sampaio, S. C., Petroni, R., Brigatte, P., Velasco, I. T., and Soriano, F. G. (2014). Phagocytic activity of LPS tolerant macrophages. *Mol. Immunol.* 60, 8–13. doi: 10.1016/j.molimm.2014.03.010

De Luca, A., Zelante, T., D'angelo, C., Zagarella, S., Fallarino, F., Spreca, A., et al. (2010). IL-22 defines a novel immune pathway of antifungal resistance. *Mucosal Immunol.* 3, 361–373. doi: 10.1038/mi.2010.22

DeAngelis, D. L., and Grimm, V. (2014). Individual-based models in ecology after four decades. *F1000Prime Rep.* 6:39. doi: 10.12703/P6-39

Dementhon, K., El-Kirat-Chatel, S., and Noël, T. (2012). Development of an *in vitro* model for the multi-parametric quantification of the cellular interactions between *Candida* yeasts and phagocytes. *PLoS ONE* 7:e32621. doi: 10.1371/journal.pone.0032621

Den Hertog, A., Van Marle, J., Van Veen, H., Van't Hof, W., Bolscher, J. G., Veerman, E. C., et al. (2005). Candidacidal effects of two antimicrobial peptides: histatin 5 causes small membrane defects, but LL-37 causes massive disruption of the cell membrane. *Biochem. J.* 388, 689–695. doi: 10.1042/BJ20042099

d'Enfert, C. (2009). Hidden killers: persistence of opportunistic fungal pathogens in the human host. *Curr. Opin. Microbiol.* 12, 358–364. doi: 10.1016/j.mib.2009.05.008

d'Ostiani, C. F., Del Sero, G., Bacci, A., Montagnoli, C., Spreca, A., Mencacci, A., et al. (2000). Dendritic cells discriminate between yeasts and hyphae of the fungus *Candida albicans* implications for initiation of T helper cell immunity *in vitro* and *in vivo*. *J. Exp. Med.* 191, 1661–1674. doi: 10.1084/jem.191.10.1661

Eyerich, S., Wagener, J., Wenzel, V., Scarponi, C., Pennino, D., Albanesi, C., et al. (2011). IL-22 and TNF-α represent a key cytokine combination for epidermal integrity during infection with *Candida albicans*. *Eur. J. Immunol.* 41, 1894–1901. doi: 10.1002/eji.201041197

Faro-Trindade, I., and Brown, G. D. (2009). "Interaction of *Candida albicans* with phagocytes," in *Phagocyte-Pathogen Interactions: Macrophages and the Host Response to Infection*, eds D. G. Russell and S. Gordon (Washington, DC: ASM Press), 437–451.

Filler, S. G. (2013). Can host receptors for fungi be targeted for treatment of fungal infections? *Trends Microbiol.* 21, 389–396. doi: 10.1016/j.tim.2013.05.006

Finkel, J. S., and Mitchell, A. P. (2011). Genetic control of *Candida albicans* biofilm development. *Nat. Rev. Microbiol.* 9, 109–118. doi: 10.1038/nrmicro2475

Folcik, V. A., An, G. C., and Orosz, C. G. (2007). The basic immune simulator: an agent-based model to study the interactions between innate and adaptive immunity. *Theor. Biol. Med. Model.* 4:39. doi: 10.1186/1742-4682-4-39

Forrest, S., and Beauchemin, C. (2007). Computer immunology. *Immunol. Rev.* 216, 176–197. doi: 10.1111/j.1600-065X.2007.00499.x

Frohner, I. E., Bourgeois, C., Yatsyk, K., Majer, O., and Kuchler, K. (2009). *Candida albicans* cell surface superoxide dismutases degrade host-derived reactive oxygen species to escape innate immune surveillance. *Mol. Microbiol.* 71, 240–252. doi: 10.1111/j.1365-2958.2008.06528.x

Ganguly, S., and Mitchell, A. P. (2011). Mucosal biofilms of *Candida albicans*. *Curr. Opin. Microbiol.* 14, 380–385. doi: 10.1016/j.mib.2011.06.001

Gow, N. A., van de Veerdonk, F. L., Brown, A. J., and Netea, M. G. (2012). *Candida albicans* morphogenesis and host defence: discriminating invasion from colonization. *Nat. Rev. Microbiol.* 10, 112–122. doi: 10.1038/nrmicro2711

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecol. Model.* 198, 115–126. doi: 10.1016/j.ecolmodel.2006.04.023

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., and Railsback, S. F. (2010). The ODD protocol: a review and first update. *Ecol. Model.* 221, 2760–2768. doi: 10.1016/j.ecolmodel.2010.08.019

Gropp, K., Schild, L., Schindler, S., Hube, B., Zipfel, P. F., and Skerka, C. (2009). The yeast *Candida albicans* evades human complement attack by secretion of aspartic proteases. *Mol. Immunol.* 47, 465–475. doi: 10.1016/j.molimm.2009.08.019

Hahn, S., Giaglis, S., Chowdury, C. S., Hösli, I., and Hasler, P. (2013). Modulation of neutrophil NETosis: interplay between infectious agents and underlying host physiology. *Semin. Immunopathol.* 35, 439–453. doi: 10.1007/s00281-013-0380-x

Hamad, M. (2012). Innate and adaptive antifungal immune responses: partners on an equal footing. *Mycoses* 55, 205–217. doi: 10.1111/j.1439-0507.2011.02078.x

Havas, K. A., Boone, R. B., Hill, A. E., and Salman, M. D. (2014). A Brucellosis disease control strategy for the Kakheti region of the Country of Georgia: an agent-based model. *Zoonoses Public Health* 61, 260–270. doi: 10.1111/zph.12066

Heinrich, R., and Schuster, S. C. (1996). *The Regulation of Cellular Systems*. New York, NY: Chapman and Hall.

Hofbauer, J., and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Horn, H., and Lackner, S. (2014). "Modeling of biofilm systems: a review," in *Productive Biofilms*, eds K. Muffler and R. Ulber (Berlin; Heidelberg: Springer-Verlag), 53–76.

Hummert, S., Bohl, K., Basanta, D., Deutsch, A., Werner, S., Theißen, G., et al. (2014). Evolutionary game theory: cells as players. *Mol. Biosyst.* 10, 3044–3065. doi: 10.1039/C3MB70602H

Hummert, S., Hummert, C., Schröter, A., Hube, B., and Schuster, S. (2010). Game theoretical modelling of survival strategies of *Candida albicans* inside macrophages. *J. Theor. Biol.* 264, 312–318. doi: 10.1016/j.jtbi.2010.01.022

Jacobsen, I. D., Wilson, D., Wächtler, B., Brunke, S., Naglik, J. R., and Hube, B. (2012). *Candida albicans* dimorphism as a therapeutic target. *Expert Rev. Anti. Infect. Ther.* 10, 85–93. doi: 10.1586/eri.11.152

Jiménez-López, C., and Lorenz, M. C. (2013). Fungal immune evasion in a model host-pathogen interaction: *Candida albicans* versus macrophages. *PLoS Pathog.* 9:e1003741. doi: 10.1371/journal.ppat.1003741

Kagami, S., Rizzo, H. L., Kurtz, S. E., Miller, L. S., and Blauvelt, A. (2010). IL-23 and IL-17a, but not IL-12 and IL-22, are required for optimal skin host defense against *Candida albicans*. *J. Immunol.* 185, 5453–5462. doi: 10.4049/jimmunol.1001153

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2011). *Systems Biology: A Textbook*. Weinheim: Wiley VCH.

Knowles, J. R. (1985). Penicillin resistance: the chemistry of β-lactamase inhibition. *Acc. Chem. Res.* 18, 97–104. doi: 10.1021/ar00112a001

Korn, T., Bettelli, E., Oukka, M., and Kuchroo, V. K. (2009). Il-17 and Th17 Cells. *Annu. Rev. Immunol.* 27, 485–517. doi: 10.1146/annurev.immunol.021908.132710

Kreft, J.-U., Plugge, C. M., Grimm, V., Prats, C., Leveau, J. H. J., Banitz, T., et al. (2013). Mighty small: observing and modeling individual microbes becomes big science. *Proc. Natl. Acad. Sci. U.S.A.* 110, 18027–18028. doi: 10.1073/pnas.1317472110

Krysan, D. J., Sutterwala, F. S., and Wellington, M. (2014). Catching fire: *Candida albicans*, macrophages, and pyroptosis. *PLoS Pathog.* 10:e1004139. doi: 10.1371/journal.ppat.1004139

Kumar, V., and Sharma, A. (2010). Neutrophils: cinderella of innate immune system. *Int. Immunopharmacol.* 10, 1325–1334. doi: 10.1016/j.intimp.2010.08.012

Kwak, M.-K., Ku, M., and Kang, S.-O. (2014). NAD^+-linked alcohol dehydrogenase 1 regulates methylglyoxal concentration in *Candida albicans*. *FEBS Lett.*, 588, 1144–1153. doi: 10.1016/j.febslet.2014.02.042

Laskowski, M., Demianyk, B. C. P., Witt, J., Mukhi, S. N., Friesen, M. R., and McLeod, R. D. (2011). Agent-based modeling of the spread of influenza-like illness in an emergency department: a simulation study. *IEEE Trans. Inf. Technol. Biomed.* 15, 877–889. doi: 10.1109/TITB.2011.2163414

Leach, M. D., Tyc, K. M., Brown, A. J. P., and Klipp, E. (2012). Modelling the regulation of thermal adaptation in *Candida albicans*, a major fungal pathogen of humans. *PLoS ONE* 7:e32467. doi: 10.1371/journal.pone.0032467

Li, R., Kumar, R., Tati, S., Puri, S., and Edgerton, M. (2013). *Candida albicans* flu1-mediated efflux of salivary histatin 5 reduces its cytosolic concentration and fungicidal activity. *Antimicrobial Agents Chemother.* 57, 1832–1839. doi: 10.1128/AAC.02295-12

Li, Y., Nguyen, M. H., Cheng, S., Schmidt, S., Zhong, L., Derendorf, H., et al. (2008). A pharmacokinetic / pharmacodynamic mathematical model accurately describes the activity of voriconazole against *Candida* spp. *in vitro*. *Int. J. Antimicrob. Agents* 31, 369–374. doi: 10.1016/j.ijantimicag.2007.11.015

Liu, P., Wu, X., Liao, C., Liu, X., Du, J., Shi, H., et al. (2014). *Escherichia coli* and *Candida albicans* induced macrophage extracellular trap-like structures with limited microbicidal activity. *PloS ONE* 9:e90042. doi: 10.1371/journal.pone.0090042

Liu, Y., Shetty, A. C., Schwartz, J. A., Bradford, L. L., Xu, W., Phan, Q. T., et al. (2015). New signaling pathways govern the host response to *C. albicans* infection in various niches. *Genome Res.* 25, 679–689. doi: 10.1101/gr.187427.114

Lopez, C. M. (2013). *The Roles of Candida albicans Gpm1p and Tef1p in Immune Evasion and Tissue Invasion of the Human Host*. Ph.D. thesis, Jena, Friedrich-Schiller-Universität Jena.

Lu, Y., Su, C., Unoje, O., and Liu, H. (2014). Quorum sensing controls hyphal initiation in *Candida albicans* through Ubr1-mediated protein degradation. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1975–1980. doi: 10.1073/pnas.1318690111

Luo, S., Skerka, C., Kurzai, O., and Zipfel, P. F. (2013). Complement and innate immune evasion strategies of the human pathogenic fungus *Candida albicans*. *Mol. Immunol.* 56, 161–169. doi: 10.1016/j.molimm.2013.05.218

Macal, C. M., North, M. J., Collier, N., Dukic, V. M., Wegener, D. T., David, M. Z., et al. (2014). Modeling the transmission of community-associated methicillin-resistant *Staphylococcus aureus*: a dynamic agent-based simulation. *J. Transl. Med.* 12:124. doi: 10.1186/1479-5876-12-124

Martin, H. L., Richardson, B. A., Nyange, P. M., Lavreys, L., Hillier, S. L., Chohan, B., et al. (1999). Vaginal Lactobacilli, microbial flora, and risk of human immunodeficiency virus type 1 and sexually transmitted disease acquisition. *J. Infect. Dis.* 180, 1863–1868. doi: 10.1086/315127

Mayer, F. L., Wilson, D., and Hube, B. (2013). *Candida albicans* pathogenicity mechanisms. *Virulence* 4, 119–128. doi: 10.4161/viru.22913

Mech, F., Thywißen, A., Guthke, R., Brakhage, A. A., and Figge, M. T. (2011). Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS ONE* 6:e19591. doi: 10.1371/journal.pone.0019591

Mech, F., Wilson, D., Lehnert, T., Hube, B., and Figge, M. T. (2014). Epithelial invasion outcompetes hypha development during *Candida albicans* infection as revealed by an image-based systems biology approach. *Cytometry A* 85A, 126–139. doi: 10.1002/cyto.a.22418

Merler, S., Ajelli, M., Fumanelli, L., Gomes, M. F. C., Piontti, A. P. Y., Rossi, L., et al. (2015). Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *Lancet Infect. Dis.* 15, 204–211. doi: 10.1016/S1473-3099(14)71074-6

Milne, G. J., Kelso, J. K., Kelly, H. A., Huband, S. T., and McVernon, J. (2008). A small community model for the transmission of infectious diseases: comparison of school closure as an intervention in individual-based models of an influenza pandemic. *PLoS ONE* 3:e4005. doi: 10.1371/journal.pone.0004005

Miramón, P., Kasper, L., and Hube, B. (2013). Thriving within the host: *Candida* spp. interactions with phagocytic cells. *Med. Microbiol. Immunol.* 202, 183–195. doi: 10.1007/s00430-013-0288-z

Moyes, D. L., and Naglik, J. R. (2011). Mucosal immunity and *Candida albicans* infection. *Clin. Dev. Immunol.* 2011:346307. doi: 10.1155/2011/346307

Naseem, M., Philippi, N., Hussain, A., Wangorsch, G., Ahmed, N., and Dandekar, T. (2012). Integrated systems view on networking by hormones in *Arabidopsis* immunity reveals multiple crosstalk for cytokinin. *Plant Cell* 24, 1793–1814. doi: 10.1105/tpc.112.098335

Netea, M. G., Brown, G. D., Kullberg, B. J., and Gow, N. A. (2008). An integrated model of the recognition of *Candida albicans* by the innate immune system. *Nat. Rev. Microbiol.* 6, 67–78. doi: 10.1038/nrmicro1815

Netea, M. G., and Maródi, L. (2010). Innate immune mechanisms for recognition and uptake of *Candida* species. *Trends Immunol.* 31, 346–353. doi: 10.1016/j.it.2010.06.007

Nobile, C. J., and Mitchell, A. P. (2006). Genetics and genomics of *Candida albicans* biofilm formation. *Cell. Microbiol.* 8, 1382–1391. doi: 10.1111/j.1462-5822.2006.00761.x

Peer, X., and An, G. (2014). Agent-based model of fecal microbial transplant effect on bile acid metabolism on suppressing *Clostridium difficile* infection: an example of agent-based modeling of intestinal bacterial infection. *J. Pharmacokinet. Pharmacodyn.* 41, 493–507. doi: 10.1007/s10928-014-9381-1

Pollmächer, J., and Figge, M. T. (2014). Agent-based model of human alveoli predicts chemotactic signaling by epithelial cells during early *Aspergillus fumigatus* infection. *PLoS ONE* 9:e111630. doi: 10.1371/journal.pone.0111630

Pruchniak, M. P., Araźna, M., and Demkow, U. (2012). Extracellular traps formation and visualization methods. *Cent. Eur. J. Immunol.* 37, 81–84.

Pumplin, N., and Voinnet, O. (2013). RNA silencing suppression by plant pathogens: defence, counter-defence and counter-counter-defence. *Nat. Rev. Microbiol.* 11, 745–760. doi: 10.1038/nrmicro3120

Qiao, Y., Liu, L., Xiong, Q., Flores, C., Wong, J., Shi, J., et al. (2013). Oomycete pathogens encode RNA silencing suppressors. *Nat. Genet.* 45, 330–333. doi: 10.1038/ng.2525

Quintin, J., Voigt, J., Voort, R., Jacobsen, I. D., Verschueren, I., Hube, B., et al. (2014). Differential role of NK cells against *Candida albicans* infection in immunocompetent or immunocompromised mice. *Eur. J. Immunol.* 44, 2405–2414. doi: 10.1002/eji.201343828

Ramage, G., Martínez, J. P., and López-Ribot, J. L. (2006). *Candida* biofilms on implanted biomaterials: a clinically significant problem. *FEMS Yeast Res.* 6, 979–986. doi: 10.1111/j.1567-1364.2006.00117.x

Ramirez-Ortiz, Z. G., and Means, T. K. (2012). The role of dendritic cells in the innate recognition of pathogenic fungi (*A. fumigatus, C. neoformans* and *C. albicans*). *Virulence* 3, 635–646. doi: 10.4161/viru.22295

Reading, C., and Cole, M. (1977). Clavulanic acid: a beta-lactamase-inhibiting beta-lactam from *Streptomyces clavuligerus*. *Antimicrobial Agents Chemother.* 11, 852–857.

Renaud, F., and De Meeus, T. (1991). A simple model of host-parasite evolutionary relationships. parasitism: compromise or conflict? *J. Theor. Biol.* 152, 319–327.

Sansregret, R., Dufour, V., Langlois, M., Daayf, F., Dunoyer, P., Voinnet, O., et al. (2013). Extreme resistance as a host counter-counter defense against viral suppression of RNA silencing. *PLoS Pathog.* 9:e1003435. doi: 10.1371/journal.ppat.1003435

Scapini, P., and Cassatella, M. A. (2014). Social networking of human neutrophils within the immune system. *Blood* 124, 710–719. doi: 10.1182/blood-2014-03-453217

Schlatter, R., Philippi, N., Wangorsch, G., Pick, R., Sawodny, O., Borner, C., et al. (2012). Integration of Boolean models exemplified on hepatocyte signal transduction. *Brief. Bioinform.* 13, 365–376. doi: 10.1093/bib/bbr065

Schuster, S., Klipp, E., and Marhl, M. (2006). "The predictive power of molecular network modelling - case studies of predictions with subsequent experimental verification," in *Discovering Biomolecular Mechanisms with Computational Biology*, ed F. Eisenhaber (Georgetown, DC: Landes Bioscience), 95–103.

Seal, J. B., Alverdy, J. C., Zaborina, O., and An, G. (2011). Agent-based dynamic knowledge representation of *Pseudomonas aeruginosa* virulence activation in the stressed gut: towards characterizing host-pathogen interactions in gut-derived sepsis. *Theor. Biol. Med. Model.* 8:33. doi: 10.1186/1742-4682-8-33

Seneviratne, C., Jin, L., and Samaranayake, L. (2008). Biofilm lifestyle of *Candida*: a mini review. *Oral Dis.* 14, 582–590. doi: 10.1111/j.1601-0825.2007.01424.x

Shankar, J., Solis, N. V., Mounaud, S., Szpakowski, S., Liu, H., Losada, L., et al. (2015). Using Bayesian modelling to investigate factors governing antibiotic-induced *Candida albicans* colonization of the GI tract. *Sci. Rep.* 5:8131. doi: 10.1038/srep08131

Siegismund, D., Schroeter, A., Lüdecke, C., Undisz, A., Jandt, K. D., Roth, M., et al. (2014a). Discrimination between random and non-random processes in early bacterial colonization on biomaterial surfaces: application of point pattern analysis. *Biofouling* 30, 1023–1033. doi: 10.1080/08927014.2014.958999

Siegismund, D., Undisz, A., Germerodt, S., Schuster, S., and Rettenmayr, M. (2014b). Quantification of the interaction between biomaterial surfaces and bacteria by 3-D modeling. *Acta Biomater.* 10, 267–275. doi: 10.1016/j.actbio.2013.09.016

Singh, A., Mahto, K. K., and Prasad, R. (2013). Lipidomics and *in vitro* azole resistance in *Candida albicans*. *OMICS* 17, 84–93. doi: 10.1089/omi.2012.0075

Smith, L. M., and May, R. C. (2013). Mechanisms of microbial escape from phagocyte killing. *Biochem. Soc. Trans.* 41, 475–490. doi: 10.1042/BST20130014

Sonnenberg, G. F., Fouser, L. A., and Artis, D. (2011). Border patrol: regulation of immunity, inflammation and tissue homeostasis at barrier surfaces by IL-22. *Nat. Immunol.* 12, 383–390. doi: 10.1038/ni.2025

Stark, H. U. (2010). Dilemmas of partial cooperation. *Evolution* 64, 2458–2465. doi: 10.1111/j.1558-5646.2010.00986.x

Stavrum, A.-K., Heiland, I., Schuster, S., Puntervoll, P., and Ziegler, M. (2013). Model of tryptophan metabolism, readily scalable using tissue-specific gene expression data. *J. Biol. Chem.* 288, 34555–34566. doi: 10.1074/jbc.M113.474908

Steele, C., and Fidel, P. L. (2002). Cytokine and chemokine production by human oral and vaginal epithelial cells in response to *Candida albicans*. *Infect. Immun.* 70, 577–583. doi: 10.1128/IAI.70.2.577-583.2002

Subbiahdoss, G., Kuijer, R., Grijpma, D. W., van der Mei, H. C., and Busscher, H. J. (2009). Microbial biofilm growth vs. tissue integration: "the race for the surface" experimentally studied. *Acta Biomater.* 5, 1399–1404. doi: 10.1016/j.actbio.2008.12.011

Swidergall, M., Ernst, A. M., and Ernst, J. F. (2013). *Candida albicans* mucin Msb2 is a broad-range protectant against antimicrobial peptides. *Antimicrob. Agents Chemother.* 57, 3917–3922. doi: 10.1128/AAC.00862-13

Szafranski-Schneider, E., Swidergall, M., Cottier, F., Tielker, D., Román, E., Pla, J., and Ernst, J. F. (2012). Msb2 shedding protects *Candida albicans* against antimicrobial peptides. *PLoS Pathog.* 8:e1002501. doi: 10.1371/journal.ppat.1002501

Tavanti, A., Campa, D., Bertozzi, A., Pardini, G., Naglik, J. R., Barale, R., et al. (2006). *Candida albicans* isolates with different genomic backgrounds display a differential response to macrophage infection. *Microbes Infect.* 8, 791–800. doi: 10.1016/j.micinf.2005.09.016

Tian, Y., Osgood, N. D., Al-Azem, A., and Hoeppner, V. H. (2013). Evaluating the effectiveness of contact tracing on tuberculosis Outcomes in Saskatchewan using individual-based modeling. *Health Educ. Behav.* 40, 98S–110S. doi: 10.1177/1090198113493910

Tierney, L., Linde, J., Müller, S., Brunke, S., Molina, J. C., Hube, B., et al. (2012). An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front. Microbiol.* 3:85. doi: 10.3389/fmicb.2012.00085

Tierney, L., Tyc, K., Klipp, E., and Kuchler, K. (2014). "Systems biology approaches to understanding and predicting fungal virulence," in *Human Fungal Pathogens, Number XII in The Mycotaed, 2nd Edn.*, ed O. Kurzai (Berlin; Heidelberg: Springer-Verlag), 45–74.

Tokarski, C., Hummert, S., Mech, F., Figge, M. T., Germerodt, S., Schroeter, A., et al. (2012). Agent-based modeling approach of immune defense against spores of opportunistic human pathogenic fungi. *Front. Microbiol.* 3:129. doi: 10.3389/fmicb.2012.00129

Tyc, K. M., and Klipp, E. (2011). Modeling dissemination of pathogenic fungi within a host : a cartoon for the interactions of two complex systems. *J. Comput. Sci. Syst. Biol.* S1:001. doi: 10.4172/jcsb.S1-001

Tyc, K. M., Kühn, C., Wilson, D., and Klipp, E. (2014). Assessing the advantage of morphological changes in *Candida albicans*: a game theoretical study. *Front. Microbiol.* 5:41. doi: 10.3389/fmicb.2014.00041

Urban, C. F., Ermert, D., Schmid, M., Abu-Abed, U., Goosmann, C., Nacken, W., et al. (2009). Neutrophil extracellular traps contain calprotectin, a cytosolic protein complex involved in host defense against *Candida albicans*. *PLoS Pathog.* 5:639. doi: 10.1371/journal.ppat.1000639

Uwamahoro, N., Verma-Gaur, J., Shen, H.-H., Qu, Y., Lewis, R., Lu, J., et al. (2014). The pathogen *Candida albicans* hijacks pyroptosis for escape from macrophages. *MBio* 5, e00003–000014. doi: 10.1128/mBio.00003-14

Uyttenhove, C., Pilotte, L., Théate, I., Stroobant, V., Colau, D., Parmentier, N., et al. (2003). Evidence for a tumoral immune resistance mechanism based

on tryptophan degradation by indoleamine 2,3-dioxygenase. *Nat. Med.* 9, 1269–1274. doi: 10.1038/nm934

van de Veerdonk, F., Joosten, L., and Netea, M. (2015). The interplay between inflammasome activation and antifungal host defense. *Immunol. Rev.* 265, 172–180. doi: 10.1111/imr.12280

Vazquez-Torres, A., Jones-Carson, J., and Balish, E. (1996). Peroxynitrite contributes to the candidacidal activity of nitric oxide-producing macrophages. *Infect. Immun.* 64, 3127–3133.

Vialas, V., Sun, Z., Loureiro y Penha, C. V., Carrascal, M., Abián, J., Monteoliva, L., et al. (2014). A *Candida albicans* peptideatlas. *J. Proteomics* 97, 62–68. doi: 10.1016/j.jprot.2013.06.020

Voigt, J. (2013). *Die Rolle von NK-Zellen in der Immunantwort Gegen Candida albicans*. Ph.D. thesis, Jena, Friedrich-Schiller-Universität Jena.

Vylkova, S., and Lorenz, M. C. (2014). Modulation of phagosomal pH by *Candida albicans* promotes hyphal morphogenesis and requires Stp2p, a regulator of amino acid transport. *PLoS Pathog.* 10:e1003995. doi: 10.1371/journal.ppat.1003995

Wellington, M., Koselny, K., Sutterwala, F. S., and Krysan, D. J. (2014). *Candida albicans* triggers NLRP3-mediated pyroptosis in macrophages. *Eukaryot. Cell* 13, 329–340. doi: 10.1128/EC.00336-13

Westerhoff, H., Riethorst, A., and Jongsma, A. (2000). Relating multidrug resistance phenotypes to the kinetic properties of their drug-efflux pumps. *Eur. J. Biochem.* 267, 5355–5368. doi: 10.1046/j.1432-1327.2000.01559.x

Whittington, A., Gow, N. A. R., and Hube, B. (2014). "From commensal to pathogen: *Candida albicans*," in *Human Fungal Pathogens number XII in The Mycota, 2nd Edn.,* ed O. Kurzai (Berlin; Heidelberg: Springer-Verlag), 3–18.

Williams, J. D. (1997). β-Lactamase inhibition and *in vitro* activity of sulbactam and sulbactam / cefoperazone. *Clin. Infect. Dis.* 24, 494–497.

Wilson, D., Thewes, S., Zakikhany, K., Fradin, C., Albrecht, A., Almeida, R., et al. (2009). Identifying infection-associated genes of *Candida albicans* in the postgenomic era. *FEMS Yeast Res.* 9, 688–700. doi: 10.1111/j.1567-1364.2009.00524.x

Yan, L., Yang, C., and Tang, J. (2013). Disruption of the intestinal mucosal barrier in *Candida albicans* infections. *Microbiol. Res.* 168, 389–395.

Yook, S.-H., Oltvai, Z. N., and Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* 4, 928–942. doi: 10.1002/pmic.200300636

Zelante, T., Iannitti, R., De Luca, A., and Romani, L. (2011). IL-22 in antifungal immunity. *Eur. J. Immunol.* 41, 270–275. doi: 10.1002/eji.201041246

Zenewicz, L. A., and Flavell, R. A. (2011). Recent advances in IL-22 biology. *Int. Immunol.* 23, 159–163. doi: 10.1093/intimm/dxr001

Zipfel, P. F., Hallström, T., and Riesbeck, K. (2013). Human complement control and complement evasion by pathogenic microbes–tipping the balance. *Mol. Immunol.* 56, 152–160. doi: 10.1016/j.molimm.2013.05.222

Zipfel, P. F., Skerka, C., Kupka, D., and Luo, S. (2011). Immune escape of the human facultative pathogenic yeast *Candida albicans*: the many faces of the *Candida* Pra1 protein. *Int. J. Med. Microbiol.* 301, 423–430. doi: 10.1016/j.ijmm.2011.04.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 April 2015
doi: 10.3389/fmicb.2015.00257

Ebola virus infection modeling and identifiability problems

Van Kinh Nguyen[1], *Sebastian C. Binder*[2], *Alessandro Boianelli*[1], *Michael Meyer-Hermann*[2,3]* *and Esteban A. Hernandez-Vargas*[1]*

[1] *Systems Medicine of Infectious Diseases, Department of Systems Immunology and Braunschweig Integrated Centre of Systems Biology, Helmholtz Centre for Infection Research, Braunschweig, Germany,* [2] *Department of Systems Immunology and Braunschweig Integrated Centre of Systems Biology, Helmholtz Centre for Infection Research, Braunschweig, Germany,* [3] *Institute for Biochemistry, Biotechnology and Bioinformatics, Technische Universität Braunschweig, Braunschweig, Germany*

OPEN ACCESS

Edited by:
Reinhard Guthke, Leibniz-Institute for Natural Product Research and Infection Biology -Hans-Knoell-Institute, Germany

Reviewed by:
Lars Kaderali, Technische Universität Dresden, Germany
Jeremie Guedj, Institut National de la Santé et de la Recherche Médicale, France

***Correspondence:**
Esteban A. Hernandez-Vargas, Systems Medicine of Infectious Diseases, Helmholtz Centre for Infection Research, Inhoffenstraße 7, 38124 Braunschweig, Germany esteban.vargas@helmholtz-hzi.de; Michael Meyer-Hermann, Department of Systems Immunology and Braunschweig Integrated Centre of Systems Biology, Helmholtz Centre for Infection Research, Inhoffenstr.7, 38124 Braunschweig, Germany mmh@theoretical-biology.de

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *28 November 2014*
Accepted: *16 March 2015*
Published: *09 April 2015*

Citation:
Nguyen VK, Binder SC, Boianelli A, Meyer-Hermann M and Hernandez-Vargas EA (2015) Ebola virus infection modeling and identifiability problems. Front. Microbiol. 6:257. doi: 10.3389/fmicb.2015.00257

The recent outbreaks of Ebola virus (EBOV) infections have underlined the impact of the virus as a major threat for human health. Due to the high biosafety classification of EBOV (level 4), basic research is very limited. Therefore, the development of new avenues of thinking to advance quantitative comprehension of the virus and its interaction with the host cells is urgently needed to tackle this lethal disease. Mathematical modeling of the EBOV dynamics can be instrumental to interpret Ebola infection kinetics on quantitative grounds. To the best of our knowledge, a mathematical modeling approach to unravel the interaction between EBOV and the host cells is still missing. In this paper, a mathematical model based on differential equations is used to represent the basic interactions between EBOV and wild-type Vero cells *in vitro*. Parameter sets that represent infectivity of pathogens are estimated for EBOV infection and compared with influenza virus infection kinetics. The average infecting time of wild-type Vero cells by EBOV is slower than in influenza infection. Simulation results suggest that the slow infecting time of EBOV could be compensated by its efficient replication. This study reveals several identifiability problems and what kind of experiments are necessary to advance the quantification of EBOV infection. A first mathematical approach of EBOV dynamics and the estimation of standard parameters in viral infections kinetics is the key contribution of this work, paving the way for future modeling works on EBOV infection.

Keywords: Ebola, mathematical modeling, kinetics, viral dynamics, identifiability, EBOV

1. Introduction

Ebola was characterized for the first time in 1976 close to the Ebola River located in the Democratic Republic of the Congo (WHO, 1978). Since then, outbreaks of EBOV among humans have appeared sporadically causing lethal diseases in several African countries, mainly in Gabon, South Sudan, Ivory Coast, Uganda, and South Africa (CDC, 2014). Among the most severe symptoms of the EBOV disease are fever, muscle pain, diarrhea, vomiting, abdominal pain and the unexplained hemorrhagic fever (Calain et al., 1999). Fatalities are predominantly associated with uncontrolled viremia and lack of an effective immune response. However, the pathogenesis of the disease is still poorly understood (Peters and Peters, 1999; Feldmann et al., 2003).

Ebola virus belongs to the family of *Filoviridae*, from Latin *filum* which means thread (Carter and Saunders, 2013). Ebola virus is classified in Tai Forest, Sudan, Zaire, Reston, and Bundibugyo. The human Ebola epidemics have been mainly related to infection by the Zaire and Sudan strains.

Filovirus virions possess several shapes, a property called pleomorphism (Feldmann et al., 2003). These shapes are appearing as either U-shaped, 6-shaped, or other configurations, e.g., **Figure 1**.

The natural hosts of EBOV still remain unsettled, but it is tenable that EBOV persists in animals which transmit the virus to non-human primates and humans (Knipe et al., 2001). It has been reported that fruit bats are capable of supporting EBOV replication without becoming ill and may serve as a major reservoir (Swanepoel et al., 1996; Knipe et al., 2001; Leroy et al., 2009; Formenty, 2014). EBOV can spread from an infected person to others through direct contact with blood or body fluids (e.g., saliva, sweat, feces, breast milk, and semen), objects (i.e., needles) that have been contaminated with the virus and infected fruit bats or primates (Peters and Peters, 1999; Feldmann et al., 2003; CDC, 2014). The 2014 Ebola epidemic is the largest ever reported in history, affecting multiple countries in West Africa and being imported to other countries: one infection case was reported in Spain while in the United States one death and two locally acquired cases in healthcare were reported (CDC, 2014).

EBOV can infect a wide variety of cell types including monocytes, macrophages, dendritic cells, endothelial cells, fibroblasts, hepatocytes, adrenal cortical cells, and several types of epithelial cells, all supporting EBOV replication. Monocytes, macrophages, and dendritic cells are early and preferred replication sites of the virus (Knipe et al., 2001). Furthermore, murine studies have revealed that EBOV can infect cells in different compartments, showing high viral titers in liver, spleen, kidney and serum (Mahanty et al., 2003).

Due to its high infectivity and fatality, the virus is classified as a biosafety level-4 agent, restricting basic research for Ebola disease (Halfmann et al., 2008). Infection parameters and quantification of the interactions between the virus and its target cells remain largely unknown. Therefore, the development of new avenues of thinking to bring forward quantitative comprehension of the relationship between the virus and the host is urgently needed. To this end, mathematical models can help to interpret experimental results on quantitative grounds. Model simulations can infer predictions to initiate further and conclusive experiments that may solve relevant scientific questions and advance knowledge of EBOV infection.

Recently, mathematical models have played a central role to capture the dynamics of different virus infections (Nowak and May, 2000). Among the most popular are HIV (Kirschner, 1996; Wu et al., 1998; Duffin and Tullis, 2002; Perelson, 2002; Hernandez-Vargas et al., 2010; Hernandez-Vargas and Middleton, 2013; Jaafoura et al., 2014), hepatitis virus (Ribeiro et al., 2002; Reluga et al., 2009; Guedj et al., 2013) and influenza virus infection models (Baccam et al., 2006; Handel et al., 2010; Smith and Perelson, 2011; Pawelek et al., 2012; Hernandez-Vargas et al., 2014). These models have been instrumental to study the mechanisms that control viral kinetics in order to provide a quantitative understanding and to formulate recommendations for treatments. Similarities of parameter values for EBOV infection to other viral infections that promote outbreaks, e.g., influenza virus infection, could be expected. Nevertheless, to the best of our knowledge, there has not been any mathematical approach until now to describe EBOV dynamics. This and the interaction of EBOV virus with non-human primate epithelial cells is the key contribution of this work.

2. Materials and Methods

2.1. Mathematical Model

The mathematical model proposed here to represent EBOV dynamics is based on the well established target cell-limited model (Nowak and May, 2000), see **Figure 2**. This has served

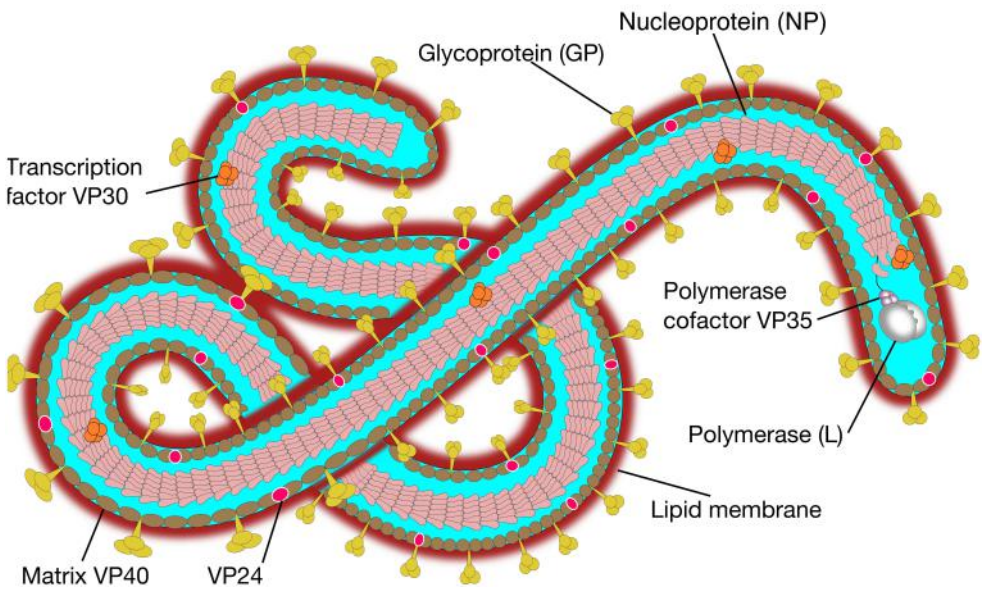


FIGURE 1 | Ebola virus molecular structure. The Ebola genome is composed of 3 leader, nucleoprotein (NP), virion protein 35 (VP35), VP40, glycoprotein (GP), VP30, VP24, polymerase (L) protein and 5 trailer (adapted from SIB SWISS Institute of Bioinformatics, 2014).

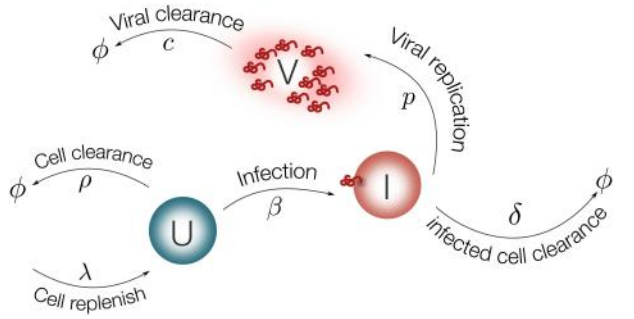


FIGURE 2 | Schematic representation of the model for EBOV infection. Target cells (U) are replenished with rate λ and die with rate ρ. Virus (V) infects target cells (U) with rate β. Infected cells are cleared with rate δ. Once cells are productively infected (I), they release virus at rate p and virus particles are cleared with rate c.

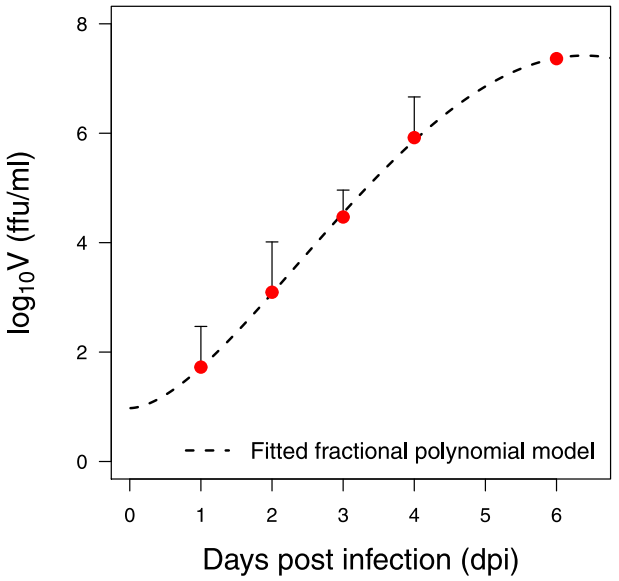


FIGURE 3 | Data preparation. Fitted statistical model for the wild-type Vero cells infected with EBOV at a *low multiplicity of infection* (MOI) (Halfmann et al., 2008)

to model several viral diseases, among them HIV infection (Wu et al., 1998; Perelson, 2002), hepatitis virus infection (Ribeiro et al., 2002) and influenza virus infection (Baccam et al., 2006; Hernandez-Vargas et al., 2014). A detailed reference for modeling of viral dynamics can be found in Nowak and May (2000).

Using ordinary differential equations (ODEs), the EBOV infection model is considered as follows:

$$\frac{dU}{dt} = \lambda - \rho U - \beta UV \tag{1}$$

$$\frac{dI}{dt} = \beta UV - \delta I \tag{2}$$

$$\frac{dV}{dt} = pI - cV \tag{3}$$

EBOV target cells can be either in a susceptible (U) or an infected state (I). Cells are replenished with a constant rate λ and die with rate ρ. Note that the condition $\lambda = U_0\rho$ should be satisfied to guarantee homeostasis in the absence of viral infection, such that only ρ is a parameter to be determined. Virus (V) infects susceptible cells with rate constant β. Infected cells are cleared with rate δ. Once cells are productively infected, they release virus at rate p and virus particles are cleared with rate c.

The initial number of susceptible cells (U_0) can be taken from the experiment in Halfmann et al. (2008) as 5×10^5. The initial value for infected cells (I_0) is set to zero. The viral titer in Halfmann et al. (2008) is measured in foci forming units per milliliter (*ffu/ml*). The initial viral load (V_0) is estimated from the data using the fractional polynomial model of second order (Royston and Altman, 1994). The best model based on the Akaike Information Criterion (AIC) is presented in **Figure 3**, providing an estimate of 9 *ffu/ml* for V_0. The parameter ρ is fixed from literature as 0.001 day^{-1} (Moehler et al., 2005). The effect of fixing this value on the model output is evaluated with a sensitivity analysis.

2.2. Experimental Data

As described in the previous section, this paper is mainly focused on the interaction between the virus and the target cells. A safe way to study the virus life cycle was proposed in Halfmann et al. (2008). The disease pathogenesis of EBOV in non-human primates is known to be more faithful in portraying the human condition than in rodents (Knipe et al., 2001). Replication kinetics of EBOV are studied in Vero cells, a cell line derived from kidney epithelial cells of African green monkeys (Halfmann et al., 2008). This non-human primate is a known source of *Filoviridae* virus infection, e.g., the European Marburg outbreak from 1967 (Knipe et al., 2001). Wild-type Vero cells and a Vero cell line expressing VP30 were tested to reveal their ability to confine EBOV to its complete replication cycle. In this study, viral kinetics for wild-type Vero cells infected with EBOV at different multiplicities of infection (MOI) were considered (Halfmann et al., 2008). The viral growth data is presented in **Figure 3**. Further details on the data, methods and experiments can be found in Halfmann et al. (2008).

2.3. Parameter Estimation

Parameter fitting is performed minimizing the root mean square (RMS) difference on log scale between the model output, $\hat{y}_i$, and the experimental measurement, y_i:

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log_{10} y_i - \log_{10} \hat{y}_i)^2} \tag{4}$$

where $n = 5$ (Halfmann et al., 2008) is the number of measurements. Differential equations are solved by R 3.1.2 (R Core Team, 2014) using the deSolve package (Soetaert et al., 2010). The minimization of RMS is performed using the Differential Evolution (DE) algorithm employing the DEoptim package (Storn, 1997; Mullen et al., 2011). The DE global optimization algorithm does not rely on initial parameter guesses and converged faster than

the other tested methods, including genetic algorithms and the quasi-Newton (BFGS, L-BFGS-B) algorithms.

2.4. Parameter Uncertainty

Viral load variability is very large for several viral infectious diseases (Mahanty et al., 2003; Baccam et al., 2006; Toapanta and Ross, 2009; Groseth et al., 2012). In order to consider the large variability of biological problems, a bootstrap method is applied to the data series presented in Halfmann et al. (2008). Bootstrapping is a statistic method for assigning measures of accuracy to estimates (Davison and Hinkley, 1997; Xue et al., 2010). The nonparametric bootstrap requires data to be independent and identically distributed while the parametric bootstrap requires to impose on the data a distribution assumption which is usually unknown. For the data in Halfmann et al. (2008), three bootstrap approaches were considered: (i) the conventional parametric approach assumes a log-normal distribution of the measurement, (ii) the nonparametric approach assumes uniform distribution in the measurement range, and (iii) the weighted bootstrap assigns to the cost function a vector of random weights from exponential distribution with mean one and variance one (Ma and Kosorok, 2005; Xue et al., 2010).

For each repetition, the model parameters are refitted to obtain the corresponding parameter distribution. The 95% confidence interval of parameter estimates is computed using the outcome of the bootstrap method (Xue et al., 2010). For each parameter, the 2.5 and 97.5% quantiles of the estimates are used to form the 95% confidence interval.

2.5. Parameter Identifiability and Sensitivity

A critical obstacle to overcome in mathematical modeling is how to verify whether model parameters are identifiable based on the measurements of output variables (Xia, 2003; Xia and Moog, 2003; Wu et al., 2008; Miao et al., 2011). A system that is algebraically identifiable may still be practically non-identifiable if the amount and quality of the measurements is insufficient and the data shows large deviations. The novel approach proposed in Raue et al. (2009) exploits the profile likelihood to determine identifiability and is considered here. This method is able to detect both structurally and practically non-identifiable parameters.

Identifiability properties are studied for the model Equations (1–3) and the data set in Halfmann et al. (2008). The idea behind this approach is to explore the parameter space for each parameter θ_i by re-optimizing the RMS with respect to all other parameters $\theta_{j\neq i}$. In particular, for each parameter θ_i, a wide range of values centered at the optimized value is generated in an adaptive manner. Re-optimization of RMS with respect to the other parameters is done for each value of parameter θ_i. The main task is to detect directions where the likelihood flattens out (Raue et al., 2009). The resulting profiles are plotted vs. each parameter range to assess the parameter identifiability visually.

In model fitting, some parameters may have little effect on the model outcome, while other parameters are so closely related that simultaneous fitting could be a difficult task. For this aspect, the scatter plots using pairs of parameters over different bootstrap replicates will be reported. Furthermore, sensitivity analysis of the estimated parameters is performed (Brun et al., 2001; Soetaert, 2014). For each data point the derivative of the corresponding modeled variable value with respect to the selected parameter is computed. The normalized sensitivity function reads as

$$\frac{\partial y_i}{\partial \Theta_j} \cdot \frac{w_{\Theta_j}}{w_{y_i}} \tag{5}$$

where y_i denotes the model variables, Θ_j is the parameter of interest, and the ratio w_{Θ_j}/w_{y_i} is the normalized factor corresponding to its nominal value (Soetaert and Petzoldt, 2010). Summary statistics of the sensitivity functions can be used to qualify the impact of the parameter on the output variables, i.e., the higher the absolute value of the sensitivity summary statistics, the more important the parameter (Brun et al., 2001). For the model in Equations (1–3), the sensitivity functions will be plotted vs. time to illustrate the parameters' role on the model output. The parameters that have little effect do not need to be fine-tuned extensively in model fitting.

2.6. Cross-Validation

It is important to prove how the model predictions will generalize to an independent data set, revealing how accurately the predictive value of a model is in practice. In this paper, the parameter set obtained from the data of wild-type Vero cells infected at low MOI is used to predict the replication kinetics of the data at high MOI presented in Halfmann et al. (2008).

3. Results

Although significant progress has been made to the identification and characterization of EBOV, human data is very limited due to the long asymptomatic periods of the virus and its high mortality. Animal models are pivotal to shed light on this lethal disease. Due to the very close similarities with the human immune system, non-human primates are the preferred animal model for several viral infections e.g., HIV). Moreover, EBOV infection has been adapted to guinea pigs and mice (Feldmann et al., 2003), serving as a flexible model in comparison to human and non-human primates. In this work, we focus on the interaction between the virus and the host cells. *In vitro* data can be very convenient due to the important simplification of the *in vivo* complexity of biological problems. Thus, for parameter fitting procedures, we consider the experimental data from Halfmann et al. (2008), which investigates EBOV kinetics in a Vero cell line.

Before rigorous optimization methods can be applied to estimate the model parameters using experimental data, the verification of parameter identifiability is required. The omission of identifiability analyses may result in incorrect fits and consequently incorrect interpretations. The identifiability analysis in the model Equations (1–3) has been broadly studied (Xia, 2003; Xia and Moog, 2003; Wu et al., 2008; Miao et al., 2011; Hernandez-Vargas et al., 2014). All parameters in the model Equations (1–3) were shown to be algebraically identifiable given measurements of viral load and initial conditions (U_0, I_0, and V_0) (Wu et al., 2008). However, the difference between structural identifiability and practical identifiability in the presence

of measurement error requires further identifiability studies. To address practical identifiability, the approach proposed by Raue et al. (2009) is considered here for the data presented in **Figure 3**.

The resulting RMS profiles in **Figure 4** for β, p and c show a convex shape of which the optimization routine can reach their minimum. Note that the profile of δ is flat in one tail, suggesting that parameter δ can be chosen arbitrarily small without affecting the fit quality (Raue et al., 2009). In spite of this, the lower bound of this parameter has a clear biological constraint. To be precise, the half-life of an infected cell cannot be longer than that of an uninfected cell. There is experimental evidence that the half-life of epithelium cells in lung is 17–18 months in average (Rawlins and Hogan, 2008). In view of this, the infected cell death rate (δ) is fixed at 10^{-3}.

Bootstrapping can provide more insights into the distribution of parameter values based on experimental data in Halfmann et al. (2008). For the sake of clarity, we present only the weighted bootstrap (Xue et al., 2010) in the results, the other two methods can be found in the supplementary material. Distributions of the model parameters are shown in **Figure 5**. Bootstrap estimates for the viral clearance (median $c = 1.05\ day^{-1}$) is slightly below other viral infection results (**Table 1**). For example, clearance of influenza virus varied from 2.6 to 15 day^{-1} in (Baccam et al., 2006; Miao et al., 2011; Pawelek et al., 2012; Hernandez-Vargas et al., 2014). This may be attributed to the fact that the viral clearance is computed for *in vitro* experiments.

EBOV is known to replicate at an unusually high rate that overwhelms the protein synthesis of infected cells (Sanchez, 2001). Consistent with this observation, bootstrap estimates revealed a very high rate of viral replication, $p = 62$ (95%CI : $31 - 580$) (**Table 1**). Although the scatter plot in **Figure 5** shows that the estimate of p can be decreased given a higher effective infection rate (β), a replication rate of at least 31.8 ffu/ml cell^{-1} day^{-1} is still needed to achieve a good fit of the viral replication kinetics in **Figure 3**.

Scatter plots are a graphical sensitivity analysis method, and a simple but useful tool to test the robustness of the results. The estimated parameters are plotted against each other. Scatter plots for the parameters in **Figure 5** provide visual evidence that these parameters strongly depend on one another such that their individual values can not be independently determined. That is, increasing the values of p increases the estimations of c. Decreasing the estimations of β increases the estimation of both c and p. However, the green curves in **Figure 5** provide the most likely

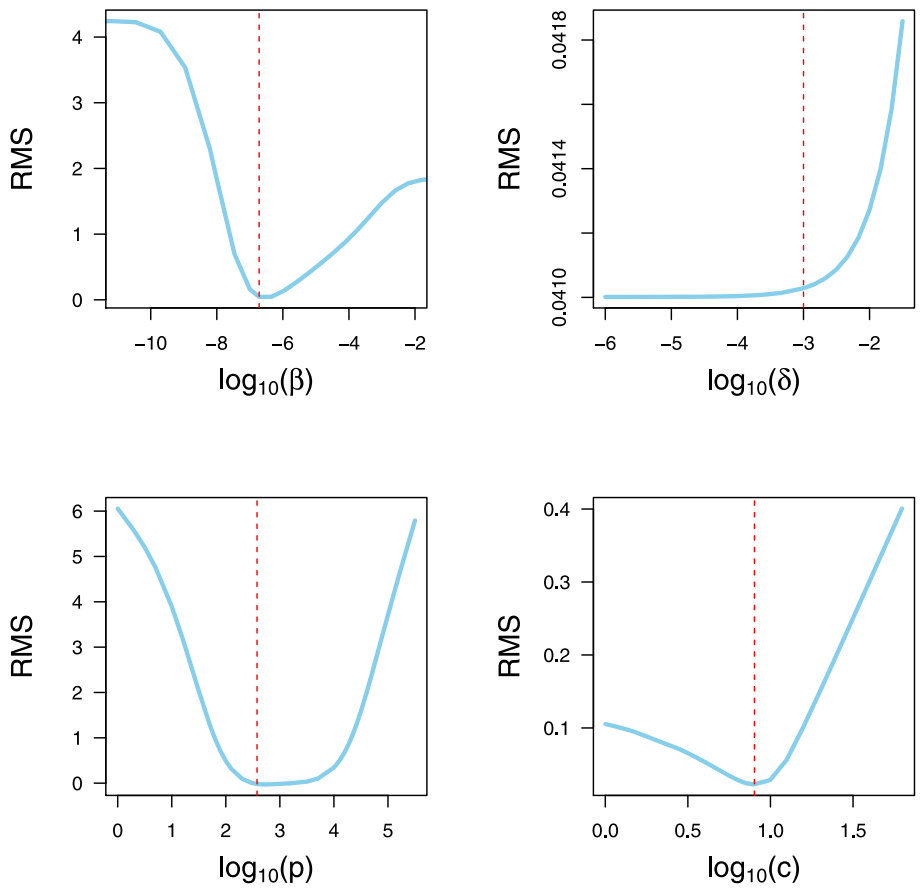


FIGURE 4 | Parameter Identifiability. RMS profile of model parameters. Each parameter is varied in a wide range around the optimized value. Subsequently, the DE algorithm is used to refit the remaining parameters to the data set of Halfmann et al. (2008). The vertical dashed lines indicate the value obtained from the optimization for all four parameters collectively.

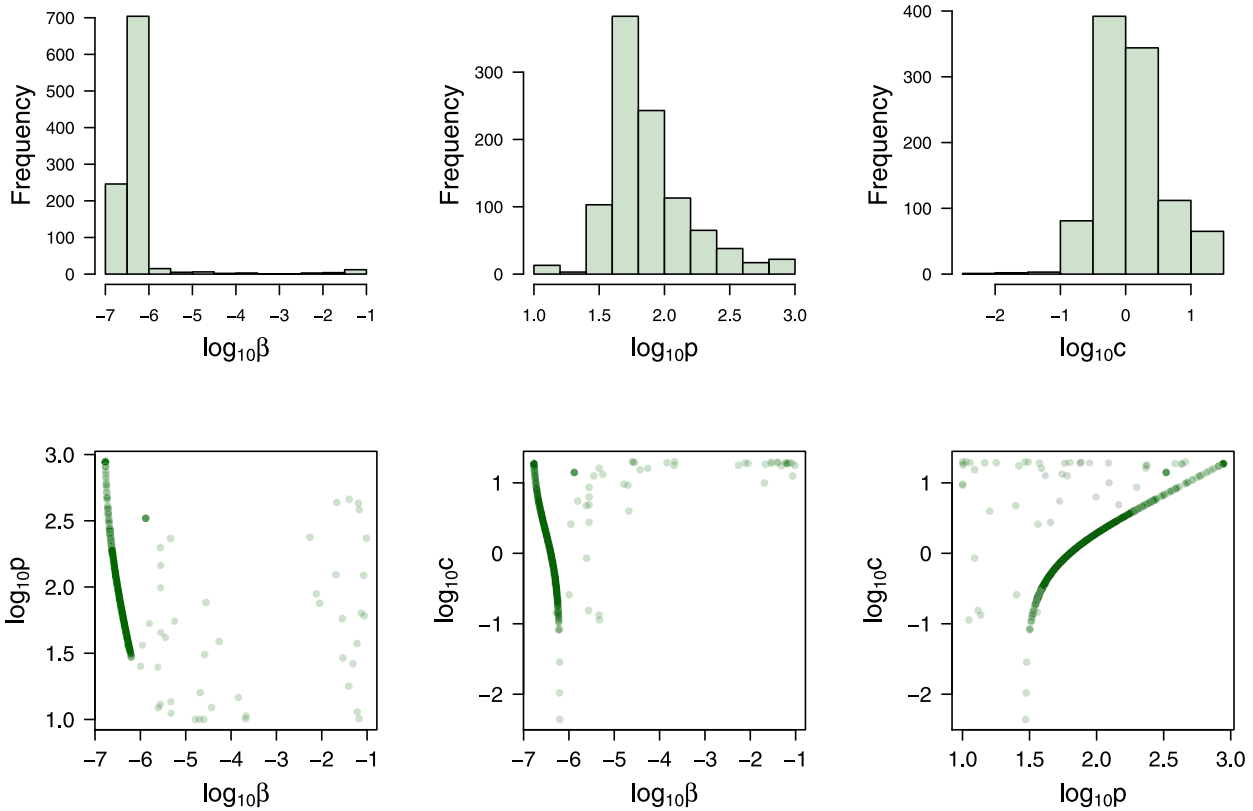


FIGURE 5 | Weighted bootstrap results. Top row: Distributions from 1000 sample estimates are presented for the three parameters: β, p and c. **Bottom row**: Scatter plot between bootstrap parameters. The parameter ρ is fixed during the bootstrapping at 0.001 (Moehler et al., 2005). Numerical values for the model Equations (1–3) are presented in the **Table 1**.

TABLE 1 | Estimates of infection parameters*.

Parameters (units)	Best fit**	Bootstrap estimates		
		2.5% quantile	Median	97.5% quantile
$\beta\left[\text{day}^{-1}(\text{ffu/ml})^{-1}10^{-7}\right]$	1.91	1.78	4.06	261.95
$p\left(\text{ffu/ml day}^{-1}\text{cell}^{-1}\right)$	378	31.80	62.91	580.69
$c\left(\text{day}^{-1}\right)$	8.02	0.18	1.05	18.76
t_{inf} (hours)	5.64	1.68	9.49	10.79

Note that these parameter should be interpreted with the discussed identifiability problems.

***Values obtained from optimization procedure to the low MOI viral titer presented in Halfmann et al. (2008).*

region where the parameters values can be found. In order to verify this intuition, we fix the viral clearance rate (c) at 4.2 (Miao et al., 2010) and then estimate the others two parameters (β and p). The results of 1000 bootstrap replicates reveal that fixing the parameter c improves the fitting with a narrow confidence interval (see Supplementary Materials 1.3).

The sensitivity study for the mathematical model Equations (1–3) is performed in a similar fashion to Brun et al. (2001); Soetaert (2014). **Figures 6A–E** show the effect on the viral load when varying the respective parameter by 10, 20 and 50% around its nominal value. It can be seen that the healthy cell death rate (ρ), which in the virus-free steady state represents the cell turnover, has little effect on the viral load kinetics. This can be attributed to the fact that the experiment was performed *in vitro* and within a short period. Similarly, the effect of the infected cell death rate (δ) can also be neglected. This could be explained by the fact that the observed Ebola viral load was not decreasing (**Figure 3**), contrary to observations in other viral infections, e.g., influenza virus (Baccam et al., 2006). The remaining three parameters (β, p, and c) are sensitive, in the sense that a small change in parameter value can lead to a large difference in viral kinetics. **Figure 6F** summarizes in detail the parameter sensitivity functions. It is clear that the three parameters β, p, and c govern the infection kinetics while the effect of the two parameters ρ and δ can be neglected for this data set. Therefore, fixing both ρ and δ is adequate for the presented problem.

Moreover, both β and p can be seen as consistently increasing the viral load because their respective sensitivity functions are

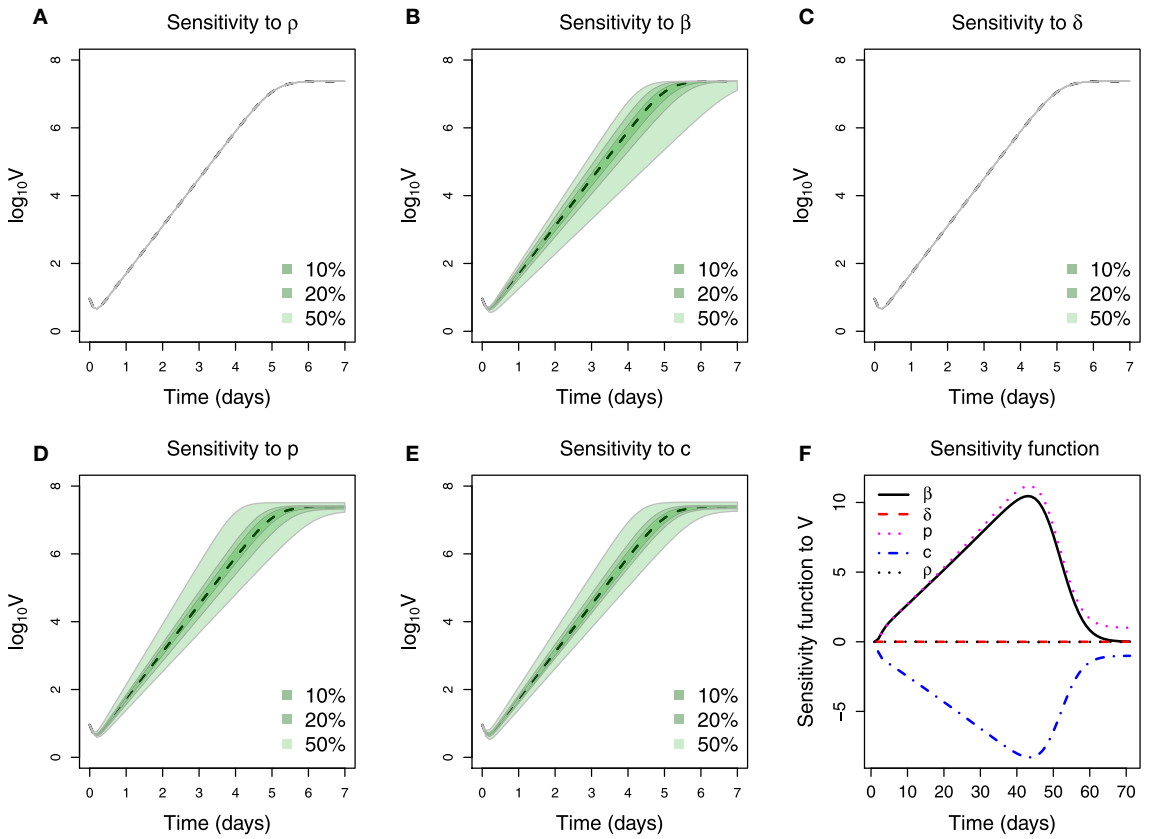


FIGURE 6 | Sensitivity of parameters. (A–E) Plotting of viral titer variation vs. time. The dashed line is the viral kinetics obtain from nominal parameter values. Three color shades in each figure represent the viral load variation range when varying the corresponding parameter by a percentage denoted in the legend. **(F)** Parameters sensitivity function over time, the values in y-axis are calculated using Equation (5).

always positive, in contrast to the parameter c. Note that the absolute magnitude of change in the sensitivity functions of these three parameters is approximately equal over time (**Figure 6F**). The strong similarity in the sensitivity functions indicates that the corresponding parameters have equivalent effect on the viral titer. For instance, the sensitivity functions of β and p are very similar so that almost the same output of viral titer will be generated by increasing β if p is decreased correspondingly. A similar statement can also be made about the relationship between c and β.

Computational simulations for the best fitting of the proposed mathematical model Equations (1–3) plotted in **Figure 7B** show that the virus grows exponentially from day 1 to 5 post infection. This is consistent with the mathematical analysis developed in Nowak et al. (1996), which deduced that the virus initially grows exponentially and can be better modeled as $\exp(r_0 t)$ while the susceptible cell population remains relatively constant, where r_0 is the leading eigenvalue which solves the equation $r_0^2 + (c + \delta)r_0 - (\beta p U_0 - c\delta) = 0$.

Viral titer peaks at high levels, more than 10^7 *ffu/ml*, which in general is 10 fold higher than those reported in influenza virus infection (Toapanta and Ross, 2009; Hernandez-Vargas et al., 2014). In addition, the viral titer reaches a plateau at day 6 and may remain at those levels (**Figure 7B**). No depletion of infected cells is observed in the period of observation. This could be a combined effect attributed to either high infection rate or high replication rate, and to the slow clearance of infected cells. To achieve virus titer levels as reported in Halfmann et al. (2008), either a high infection rate (β) of susceptible cells, or a high replication rate is required (**Figure 5**). Note that even though these estimations were performed *in vitro*, *in vivo* murine studies for EBOV infection (Mahanty et al., 2003) showed similar kinetics and time scales as those presented in **Figure 7B**.

3.1. Transmission Measures

Infectivity is a critical parameter to assess the ability of a pathogen to establish an infection (Diekmann et al., 1990). To determine infectivity, we compute the reproductive number (R_0), which is defined as the expected number of secondary infections produced by an infected cell in its lifetime (Diekmann et al., 1990; Heffernan et al., 2005). On the one hand, if R_0 is less than one, each infected individual produces on average less than one infected individual, and therefore the infection will be cleared from the population. On the other hand, if R_0 is greater than one, the pathogen is able to invade the susceptible population. This epidemiological concept can be applied to the model Equations

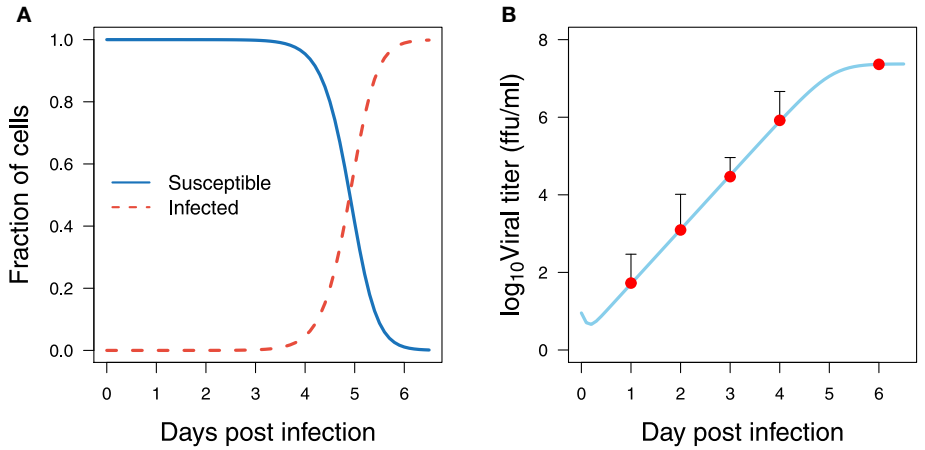


FIGURE 7 | Model fitting for EBOV kinetics. Viral titer data with low MOI from Halfmann et al. (2008) and simulations from the best fit shown in **Table 1** are in panel **(A)** for the host cells and **(B)** for the viral titer.

(1–3) and computed as follows (Nowak et al., 1996):

$$R_0 = \frac{\lambda p \beta}{c \rho \delta} \tag{6}$$

As expected, the estimated reproductive number in EBOV infection is very high, see **Figure 8A** and numerical results in **Table 1**. These results can be attributed to the fact that no depletion of virus was observed and to a slow clearance of infected cells. Thus, both parameters δ and c increase the value of R_0. Note that very high estimates of the reproductive number in highly viremic influenza virus strains from *in vitro* experiments have also been reported, with an average of 1.7×10^3 (Pinilla et al., 2012). It is worth to mention that fitting the model to *in vitro* data in Halfmann et al. (2008) could lead to small estimates for c and δ in comparison to an *in vivo* situation. Nevertheless, estimates of the epithelial cell half-life were 6 months in the trachea and 17 months in the lungs in average (Bowden, 1983; Rawlins and Hogan, 2008), which corresponds to a δ equal to 0.003 and 0.001, respectively. As mentioned previously, the δ was fixed at 0.001 in the computation of R_0. Therefore, the estimated values of R_0 interval are very likely to be positioned in a biologically plausible range, especially the upper bound. Notwithstanding, the estimate of R_0 presented here should be interpreted with care within the limits of the data used.

Recent viral modeling works (Holder et al., 2011; Pinilla et al., 2012) have also introduced the term *infecting time*, which represents the amount of time required for a single infectious cell to cause the infection of one more cell within a completely susceptible population. Strains with a shorter infecting time have a higher infectivity (Holder et al., 2011; Pinilla et al., 2012). From model Equations (1–3), this measure can be computed as follows:

$$t_{\text{inf}} = \sqrt{\frac{2}{p \beta U_0}} \tag{7}$$

Bootstrap results showed that EBOV possesses an average infecting times of 9.49 h (**Table 1**) which is approximately 7 times

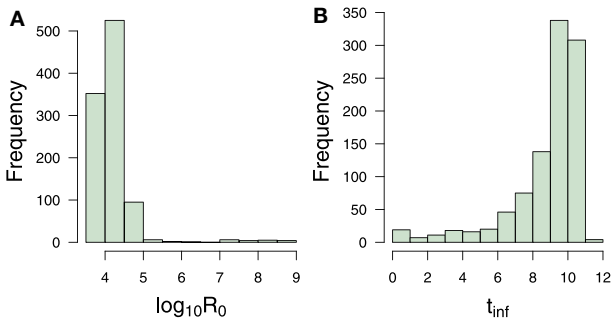


FIGURE 8 | Transmission measures. Bootstrap estimate of **(A)** reproductive number and **(B)** *infecting time* in hours. Numerical values can be found in **Table 1**.

slower than the infecting time of influenza virus (Holder et al., 2011). This number provides a reasonable explanation for the kinetics of susceptible cells which slowly decrease from day 1 to day 4 (**Figure 7A**), and quickly deplete within the last 2 days. This number could also explain the absence of viral replication within the first 5.6 h after infection. This period corresponds to the short decreasing period observed in **Figure 7B**. The initial decrease of viral load thus can be attributed to self-clearance of the virus when some viruses have infected cells but are not yet able to replicate.

The infectivity parameters in **Figure 8** characterize the EBOV infection kinetics in the data in Halfmann et al. (2008). The slow infection time of EBOV is compensated by its efficient replication. As a result, a short delay is followed by a massive amount of virus. The above infectivity parameters contributed an explanation for the high levels of viral load even when the susceptible cells were already depleted at the end of the experiment.

The best set of estimated parameters is challenged to validate the data at high multiplicities of infection (MOI) in Halfmann et al. (2008). The initial viral load is estimated using the fractional polynomial model of second order providing V_0 at 460 ffu/ml.

Figure 9 shows that the parameters derived from data at low multiplicity of infection are still consistent with data generated at high multiplicity of infection. The predicted kinetics follows the experimental data closely when changing the initial condition of the viral titer to 50 folds higher.

4. Discussion

Ebola virus (EBOV) is highly pathogenic for humans, being nowadays one of the most lethal pathogens worldwide. Ebola fatalities are predominantly associated with uncontrolled viremia and lack of an effective immune response (i.e., low levels of antibodies and no cellular infiltrates at sites of infection) (Feldmann et al., 2003).

The work presented here focused on the interaction between EBOV and the host cells, i.e., epithelial cells of green monkey. Experimental data on the Vero cell line from non-human primates could help to better understand the virus infection dynamics in humans (Knipe et al., 2001). However, the *in vitro* studies must be translated carefully to avoid over-interpretation to the *in vivo* context, which can sometimes lead to erroneous conclusions. Especially, the EBOV infection has been known to have abnormal behavior *in vivo* where different cells types and the immune system are involved (Knipe et al., 2001). Additionally, given the fact that EBOV exhibits an asymptomatic period in humans (Leroy et al., 2000), the viral dynamics model *in vivo* should take the eclipse phase into consideration. This feature can be modeled by adding an appropriate eclipse phase term as has been done previously (Moehler et al., 2005; Baccam et al., 2006). Nevertheless, given the problem of parameter identifiability exposed in the results, a complex model would not bring any better understanding. Once more data would become available, future work could attempt to address this issue, especially in the *in vivo* context.

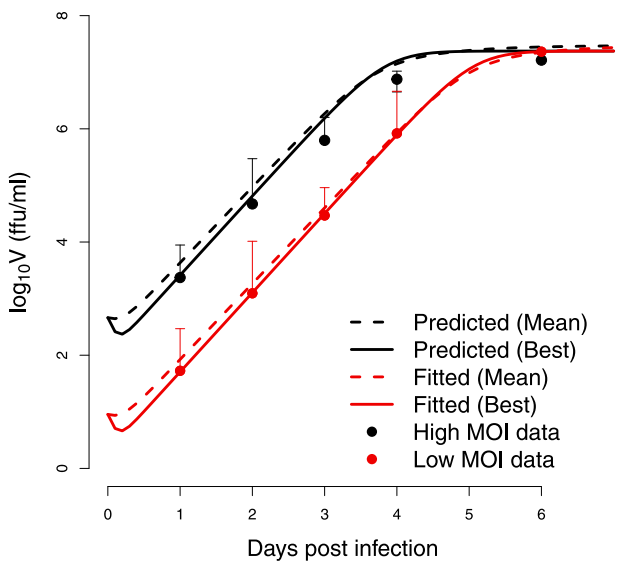


FIGURE 9 | Cross-validation. Test of estimated parameters on an independent set of data. The viral replication kinetics in wild-type Vero cells infected with EBOV at a high multiplicities of infection (MOI) in Halfmann et al. (2008) are modeled starting from a higher initial viral load of $V_0 = 460$ ffu/ml. The *(Mean)* indicates the predicted kinetics using parameters obtained from bootstrap while *(Best)* refers to the predicted kinetics using the parameters resulting from the optimization.

The exposed identifiability issues in the results reveal the problematic of parameter estimation using solemnly the viral load measurements. Here, our efforts to cope thoroughly with the identifiability issues spotted the current restrictions on the estimated parameters. These restrictions cannot be resolved without the progress of new experiments, more measurements are necessary to sort out the identifiability problems presented here, e.g., measurements of infected and non-infected cells. Another possible experiment is to determine the EBOV clearance rate in the absence of target cells. For instance, Pinilla et al. (2012) employed an experiment in a similar fashion to determine the viral infectivity loss (c). Known influenza virus titers were incubated without target cells and followed up to determine the remaining infectious titers (Pinilla et al., 2012). In this way the approximate values of the viral clearance rate could be determined and provide a more accurate estimates for the whole set of kinetics parameters, as shown in the Supplemental Material 1.3.

The high EBOV replication reported here is in agreement with recent findings by Misasi and Sullivan (2014) as well as documented in Knipe et al. (2001), reporting that early and coordinated disruptions by Ebola genes and proteins (VP24, VP30, and VP35) lead to elevated levels of virus replication. The bootstrap results suggested that the EBOV average infecting time is approximately 9.5 h, at least 5 fold slower than estimations from influenza virus infection (Pinilla et al., 2012). These simulations outline the EBOV kinetics in the data from Halfmann et al. (2008), suggesting that a slow infecting time of EBOV is compensated by its efficient replication.

The model results suggested that the saturation of viral growth as observed in the data is induced by the loss of susceptible cells. This result has to be re-evaluated with a more complete data set, as the present data set would also be appropriately described by a logistic-growth model (data not shown) with an unspecific limitation of resources. However, a logistic model can explain only the growth behavior of the virus. As pointed out before (Wu et al., 2008), a higher resolution of the data and later time points which exhibit the long-term behavior of the viral load are required for a full determination of the mechanisms at work.

EBOV infection from *in vitro* and even murine systems may differ considerably from humans. The latency phase in human is much longer than in animals and EBOV symptoms in humans may appear from 2 to 21 days after exposure to the virus, having an average time of 8–10 days (Peters and Peters, 1999). Remarkably, mice infected by intra-peritoneal injection develop symptomatic infection where EBOV will increase rapidly at day 4 and continue to increase until day 6, with death occurring at day 6–7 post-infection (Mahanty et al., 2003). These experimental observations are compatible with our simulation results, suggesting that the growth of infected cells starts at day 3 post infection (**Figure 7**) while almost the whole susceptible cell pool is depleted at day 6 post infection. It is worth to mention that EBOV kinetics were similar in different tissue compartments (Mahanty et al., 2003): liver, spleen, kidney and serum. Consequently,

further modeling approaches should address the EBOV kinetics in different compartments of the infected host.

The *in vitro* system may mimick a human context where the immune response against EBOV is not working adequately. The onset of a CD8+ T cell response as well as of the antibody response (Gupta et al., 2003) rely on early regulation of cytokines in the asymptomatic phase of the disease (Mahanty et al., 2003; Ebihara et al., 2006; García-Sastre and Biron, 2006). Human EBOV infection revealed that patients infected by the Sudan strain had lower levels of tumor necrosis factor TNF-α and interferon IFN-γ compared to those found in patients with fatal Zaire strain infection (Hutchinson and Rollin, 2007). Additionally, the levels of IFN-α were found significantly higher in surviving patients with Sudan strain infection (Hutchinson and Rollin, 2007), whereas the levels of IL-6, IL-8, IL-10, and macrophage inflammatory proteins were higher in patients with fatal infections (Hutchinson and Rollin, 2007). Therefore, modeling the effects of IFN-I would limit the number of infected cells by the introduction of a resistant state with a possible impact on the value of the viral replication rate (p). Future modeling studies need to quantify the situation *in vivo* where the effect of the immune system is taken into account.

The modeling work developed in this paper paves the way for future mathematical models and experiments to shed light on the reasons for less efficient control of Ebola virus infections. Determining empirically the EBOV clearance rate in the absence of target cells would fulfill the picture of EBOV kinetics *in vitro*. In addition, due to the critical relevance of the cytokine effects in EBOV pathogenesis, future modeling attempts should be directed to establish a more detailed model of interactions between the relevant cytokines and EBOV. Further insights into immunology and pathogenesis of EBOV will help to improve the outcome of this lethal disease.

Acknowledgments

This work was supported by iMed—the Helmholtz Initiative on Personalized Medicine. In addition, we thank the support provided by the Measures for the Establishment of Systems Medicine (e:Med) projects in Systems Immunology and Image Mining in Translational Bio- marker Research (SYSIMIT) and in identification of predictive response and resistance factors to targeted therapy in gastric cancer using a systems medicine approach (SYS-Stomach) by the Federal Ministry of Education and Research (BMBF), Germany. VKN has been supported by the President's Initiative and Networking Funds of the Helmholtz Association of German Research Centres (HGF) under contract number VH-GS-202.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00257/abstract

References

Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G., and Perelson, A. S. (2006). Kinetics of influenza A virus infection in humans. *J. Virol.* 80, 7590–7599. doi: 10.1128/JVI.01623-05

Bowden, D. (1983). Cell turnover in the lung. *Am. Rev. Res. Dis.* 128(2 Pt 2), S46–S48.

Brun, R., Reichert, P., and Kuensch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Res. Res.* 37, 1015–1030. doi: 10.1029/2000WR900350

Calain, P., Bwaka, M. A., Colebunders, R., Roo, A. D., Guimard, Y., Katwiki, K. R., et al (1999). Ebola hemorrhagic fever in kikwit , democratic republic of the congo : clinical Observations in 103 Patients. *J. Infect. Dis.* 179(Suppl. 1), 1–7. doi: 10.1086/514308

Carter, J., and Saunders, V. (2013). *Virology: Principle and Applications*. Chichester, UK: John Wiley.

CDC. (2014). *CDC Report to Ebola Virus Disease 2014*. Technical report.

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University.

Diekmann, O., Heesterbeek, J., and Metz, J. (1990). On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases. *J. Math. Biol.* 28, 365–382. doi: 10.1007/BF00178324

Duffin, R. P., and Tullis, R. H. (2002). Mathematical models of the complete course of HIV infection and AIDS. *J. Theor. Med.* 4, 215–221. doi: 10.1080/1027366021000051772

Ebihara, H., Takada, A., Kobasa, D., Jones, S., Neumann, G., Theriault, S., et al. (2006). Molecular determinants of Ebola virus virulence in mice. *PLoS Pathog.* 2:e73. doi: 10.1371/journal.ppat.0020073

Feldmann, H., Jones, S., Klenk, H.-d., and Schnittler, H.-j. (2003). Ebola virus: from discovery to vaccine. *Nat. Rev.* 3, 677–685. doi: 10.1038/nri1154

Formenty, P. (2014). "Chapter 9 - ebola virus disease," in *Emerging Infectious Diseases*, eds N. Ergnl, F. Can, L. Madoff, and M. Akova (Amsterdam: Academic Press), 121–134.

García-Sastre, A., and Biron, C. a. (2006). Type 1 interferons and the virus-host relationship: a lesson in détente. *Science* 312, 879–882. doi: 10.1126/science.1125676

Groseth, A., Marzi, A., Hoenen, T., Herwig, A., Gardner, D., Becker, S., et al. (2012). The Ebola virus glycoprotein contributes to but is not sufficient for virulence *in vivo*. *PLoS Pathog.* 8:e1002847. doi: 10.1371/journal.ppat.1002847

Guedj, J., Dahari, H., Rong, L., Sansone, N. D., Nettles, R. E., Cotler, S. J., et al. (2013). Modeling shows that the NS5A inhibitor daclatasvir has two modes of action and yields a shorter estimate of the hepatitis C virus half-life. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3–8. doi: 10.1073/pnas.1203110110

Gupta, M., Mahanty, S., Greer, P., Towner, J. S., Shieh, W.-J., Zaki, S. R., et al. (2003). Persistent Infection with Ebola Virus under Conditions of Partial Immunity. *J. Virol.* 78, 958–967. doi: 10.1128/JVI.78.2.958-967.2004

Halfmann, P., Kim, J., Ebihara, H., Noda, T., Neumann, G., Feldmann, H., et al. (2008). Generation of biologically contained Ebola viruses. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1129–1133. doi: 10.1073/pnas.0708057105

Handel, A., Longini, I. M., and Antia, R. (2010). Towards a quantitative understanding of the within-host dynamics of influenza A infections. *J. R. Soc. Interf.* 7, 35–47. doi: 10.1098/rsif.2009.0067

Heffernan, J. M., Smith, R. J., and Wahl, L. M. (2005). Perspectives on the basic reproductive ratio. *J. R. Soc. Interf.* 2, 281–293. doi: 10.1098/rsif.2005.0042

Hernandez-Vargas, E. A., Colaneri, P., Middleton, R. H., and Blanchini, F. (2010). Discrete-time control for switched positive systems with application to mitigating viral escape. *Int. J. Rob. Nonlin. Control* 21, 1093–1111. doi: 10.1002/rnc.1628

Hernandez-Vargas, E. A., and Middleton, R. H. (2013). Modeling the three stages in HIV infection. *J. Theor. Biol.* 320, 33–40. doi: 10.1016/j.jtbi.2012.11.028

Hernandez-Vargas, E. a., Wilk, E., Canini, L., Toapanta, F. R., Binder, S. C., Uvarovskii, A., et al. (2014). Effects of aging on influenza virus infection dynamics. *J. Virol.* 88, 4123–4131. doi: 10.1128/JVI.03644-13

Holder, B. P., Simon, P., Liao, L. E., Abed, Y., Bouhy, X., Beauchemin, C. A. A., et al. (2011). Assessing the *in vitro* fitness of an oseltamivir-resistant seasonal A/H1N1 influenza strain using a mathematical model. *PLoS ONE* 6:e14767. doi: 10.1371/journal.pone.0014767

Hutchinson, K. L., and Rollin, P. E. (2007). Cytokine and chemokine expression in humans infected with Sudan Ebola virus. *J. Infect. Dis.* 196(Suppl.), S357–S363. doi: 10.1086/520611

Jaafoura, S., de Goër de Herve, M. G., Hernandez-Vargas, E. A., Hendel-Chavez, H., Abdoh, M., Mateo, M. C., et al. (2014). Progressive contraction of the latent HIV reservoir around a core of less-differentiated CD4(+) memory T Cells. *Nat. Commun.* 5, 1–8. doi: 10.1038/ncomms6407

Kirschner, D. (1996). Using mathematics to understand HIV immune dynamics. *AMS Notices* 43, 191–202.

Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B., et al. (2001). *Field Virology*. Philadelphia, PA: Lippincott Williams and Wilkins.

Leroy, E. M., Baize, S., Volchkov, V. E., Fisher-Hoch, S. P., Georges-Courbot, M. C., et al. (2000). Human asymptomatic Ebola infection and strong inflammatory response. *Lancet* 355, 2210–2215. doi: 10.1016/S0140-6736(00)02405-3

Leroy, E. M., Epelboin, A., Mondonge, V., Pourrut, X., Gonzalez, J. P., Muyembe-Tamfum, J. J., et al. (2009). Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector Borne Zoonotic Dis.* 9, 723–728. doi: 10.1089/vbz.2008.0167

Ma, S., and Kosorok, M. R. (2005). Robust semiparametric m-estimation and the weighted bootstrap. *J. Multivar. Anal.* 96, 190–217. doi: 10.1016/j.jmva.2004.09.008

Mahanty, S., Gupta, M., Paragas, J., and Bray, M. (2003). Protection from lethal infection is determined by innate immune responses in a mouse model of Ebola virus infection. *Virology* 312, 415–424. doi: 10.1016/S0042-6822(03)00233-2

Miao, H., Hollenbaugh, J. A., Zand, M. S., Holden-Wiltse, J., Mosmann, T. R., Perelson, A. S., et al. (2010). Quantifying the early immune response and adaptive immune response kinetics in mice infected with influenza A virus. *J. Virol.* 84, 6687–6698. doi: 10.1128/JVI.00266-10

Miao, H., Xia, X., Perelson, A. S., and Wu, H. (2011). on Identifiability of Nonlinear Ode Models and Applications in Viral Dynamics. *SIAM Rev. Soc. Industr. Appl. Math.* 53, 3–39. doi: 10.1137/090757009

Misasi, J., and Sullivan, N. J. (2014). Camouflage and misdirection: The full-on assault of ebola virus disease. *Cell* 159, 477–486. doi: 10.1016/j.cell.2014.10.006

Moehler, L., Flockerzi, D., Sann, H., and Reichl, U. (2005). Mathematical model of influenza a virus production in large-scale microcarrier culture. *Biotechnol. Bioeng.* 90, 46–58. doi: 10.1002/bit.20363

Mullen, K., Ardia, D., Gil, D., Windover, D., and Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *J. Stat. Softw.* 40, 1–26. Available online at: http://www.jstatsoft.org/v40/i06/bibtex

Nowak, M. A., Bonhoeffer, S., Hill, A. M., Boehme, R., Thomas, H. C., and McDade, H. (1996). Viral dynamics in hepatitis b virus infection. *Proc. Natl. Acad. Sci. U.S.A.* 93, 4398–4402. doi: 10.1073/pnas.93.9.4398

Nowak, M. A., and May, R. (2000). *Virus Dynamics: Mathematical Principles of Immunology and Virology*, *Vol. 291*. Oxford, UK: Oxford University Press.

Pawelek, K. A., Huynh, G. T., Quinlivan, M., Cullinane, A., Rong, L., and Perelson, A. S. (2012). Modeling within-host dynamics of influenza virus infection including immune responses. *PLoS Comp. Biol.* 8:e1002588. doi: 10.1371/journal.pcbi.1002588

Perelson, A. S. (2002). Modelling viral and immune system dynamics. *Nat. Rev. Immunol.* 2, 28–36. doi: 10.1038/nri700

Peters, C., and Peters, J. (1999). An introduction to Ebola: the virus and the disease. *J. Infect. Dis.* 179, ix–xvi. doi: 10.1086/514322

Pinilla, L. T., Holder, B. P., Abed, Y., Boivin, G., and Beauchemin, C. A. (2012). The H275Y neuraminidase mutation of the pandemic A/H1N1 influenza virus lengthens the eclipse phase and reduces viral output of infected cells, potentially compromising fitness in ferrets. *J. Virol.* 86, 10651–10660. doi: 10.1128/JVI.07244-11

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raue, a., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* 25, 1923–1929. doi: 10.1093/bioinformatics/btp358

Rawlins, E. L., and Hogan, B. L. M. (2008). Ciliated epithelial cell lifespan in the mouse trachea and lung. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 295, L231–L234. doi: 10.1152/ajplung.90209.2008

Reluga, T., Dahari, H., and Perelson, A. (2009). Analysis of hepatitis C virus infection models with hepatocyte homeostasis. *SIAM J. Appl. Math.* 69, 999–1023. doi: 10.1137/080714579

Ribeiro, R. M., Lo, A., and Perelson, A. S. (2002). Dynamics of hepatitis B virus infection. *Microbes Infect.* 4, 829–835. doi: 10.1016/S1286-4579(02)01603-9

Royston, P., (Imperial College School of Medicine, London) and Altman, D. G., (Imperial Cancer Research Fund, London). (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl. Statist.* 43, 429–467.

Sanchez, A. (2001). *Filoviridae: Marburg and Ebola Viruses*. Philadelphia, PA: John Wiley & Sons, Ltd.

SIB SWISS Institute of Bioinformatics (2014). *Ebolavirus.* Genève.

Smith, A. M., and Perelson, A. S. (2011). Influenza A virus infection kinetics: quantitative data and models. *Syst. Biol. Med.* 3, 429–445. doi: 10.1002/wsbm.129

Soetaert, K. (2014). Package rootsolve: roots, gradients and steady-states in r.

Soetaert, K., and Petzoldt, T. (2010). Inverse modelling, sensitivity and monte carlo analysis in R using package FME. *J. Stat. Softw.* 33, 1–28. Available online at: http://www.jstatsoft.org/v33/i03/bibtex

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving differential equations in r: Package desolve. *J. Stat. Softw.* 33, 1–25. Available online at: http://www.jstatsoft.org/v33/i09/bibtex

Storn, R. (1997). Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optimizat.* 11, 341–359. doi: 10.1023/A:1008202821328

Swanepoel, R., Leman, P. A., Burt, F. J., Zachariades, N. A., Braack, L. E., Ksiazek, T. G., et al. (1996). Experimental inoculation of plants and animals with Ebola virus. *Emerg. Infect. Dis.* 2, 321–325. doi: 10.3201/eid0204.960407

Toapanta, F. R., and Ross, T. M. (2009). Impaired immune responses in the lungs of aged mice following influenza infection. *Res. Res.* 10, 1–19. doi: 10.1186/1465-9921-10-112

WHO (1978). *Ebola Haemorrhagic Fever in Sudan, 1976: Report of a World Health Organization International Study Team*. Technical report.

Wu, H., Ding, A. A., and De Gruttola, V. (1998). Estimation of HIV dynamic parameters. *Stat. Med.* 17, 2463–2485. doi: 10.1002/(SICI)1097-0258(19981115)17:21<2463::AID-SIM939>3.0.CO;2-A

Wu, H., Zhu, H., Miao, H., and Perelson, A. S. (2008). Parameter identifiability and estimation of HIV/AIDS dynamic models. *Bull. Math Biol.* 70, 785–799. doi: 10.1007/s11538-007-9279-9

Xia, X. (2003). Estimation of HIV/AIDS parameters. *Automatica* 39, 1983–1988. doi: 10.1016/S0005-1098(03)00220-6

Xia, X., and Moog, C. (2003). Identifiability of nonlinear systems with application to HIV/AIDS models. *IEEE Trans. Automat. Control* 48, 330–336. doi: 10.1109/TAC.2002.808494

Xue, H., Miao, H., and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Ann. Statist.* 38, 2351–2387. doi: 10.1214/09-AOS784

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 March 2015
doi: 10.3389/fmicb.2015.00171

Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study

Andreas Dix[1], *Kerstin Hünniger*[2], *Michael Weber*[2], *Reinhard Guthke*[1], *Oliver Kurzai*[2] *and Jörg Linde*[1]*

[1] Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Jena, Germany, [2] Septomics Research Centre, Friedrich Schiller University and Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Jena, Germany

OPEN ACCESS

Edited by:
Tunahan Cakir,
Gebze Technical University, Turkey

Reviewed by:
Pinar Pir,
Babraham Institute, UK
Daniel Vis,
Netherlands Cancer Institute, Netherlands

***Correspondence:**
Jörg Linde, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute, Beutenbergstraße 11a, Jena 07745, Germany
joerg.linde@hki-jena.de

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *03 November 2014*
Accepted: *15 February 2015*
Published: *11 March 2015*

Citation:
Dix A, Hünniger K, Weber M, Guthke R, Kurzai O and Linde J (2015) Biomarker-based classification of bacterial and fungal whole-blood infections in a genome-wide expression study. Front. Microbiol. 6:171. doi: 10.3389/fmicb.2015.00171

Sepsis is a clinical syndrome that can be caused by bacteria or fungi. Early knowledge on the nature of the causative agent is a prerequisite for targeted anti-microbial therapy. Besides currently used detection methods like blood culture and PCR-based assays, the analysis of the transcriptional response of the host to infecting organisms holds great promise. In this study, we aim to examine the transcriptional footprint of infections caused by the bacterial pathogens *Staphylococcus aureus* and *Escherichia coli* and the fungal pathogens *Candida albicans* and *Aspergillus fumigatus* in a human whole-blood model. Moreover, we use the expression information to build a random forest classifier to classify if a sample contains a bacterial, fungal, or mock-infection. After normalizing the transcription intensities using stably expressed reference genes, we filtered the gene set for biomarkers of bacterial or fungal blood infections. This selection is based on differential expression and an additional gene relevance measure. In this way, we identified 38 biomarker genes, including *IL6*, *SOCS3*, and *IRG1* which were already associated to sepsis by other studies. Using these genes, we trained the classifier and assessed its performance. It yielded a 96% accuracy (sensitivities >93%, specificities >97%) for a 10-fold stratified cross-validation and a 92% accuracy (sensitivities and specificities >83%) for an additional test dataset comprising *Cryptococcus neoformans* infections. Furthermore, the classifier is robust to Gaussian noise, indicating correct class predictions on datasets of new species. In conclusion, this genome-wide approach demonstrates an effective feature selection process in combination with the construction of a well-performing classification model. Further analyses of genes with pathogen-dependent expression patterns can provide insights into the systemic host responses, which may lead to new anti-microbial therapeutic advances.

Keywords: immune response, microarray, feature selection, systems biology, decision tree based methods, fungal pathogens

1. Introduction

Sepsis is a critical medical condition with high mortality rates. It is characterized by a dysregulation of the inflammatory response of the host due to a microbial infection. The uncontrolled inflammation can lead to tissue and organ damage, eventually resulting in death of the patient (Rittirsch et al., 2008). The incidence of sepsis has been increasing worldwide (Engel et al., 2007; Martin, 2012). In fact, sepsis is the 10th most common cause of death with a mortality rate of 20–50% in the US (Martin et al., 2003). The most frequent causative pathogens are bacteria, most commonly staphylococci and Enterobacteriaceae like *E. coli* (Martin, 2012). While the overall incidence of sepsis is increasing about 5–10% every year, the cases of sepsis caused by fungi have increased by more than 200% in the US between 1979 and 2000 (Martin et al., 2003). Since both types of pathogens, bacteria and fungi, require fundamentally different anti-microbial therapies, the early classification is crucial. Furthermore, it has been shown that prompt treatment is a prerequisite for successful therapy, as each hour of delay reduces the chances of survival on average by 8% (Kumar et al., 2006). This direct relation emphasizes the necessity for quick and reliable classification methods.

Blood cultures (BCs) and PCR-based assays are currently the standard diagnosis techniques to detect causative pathogens. While BCs aim for the isolation, identification, and susceptibility tests of microorganisms (Westh et al., 2009), molecular pathogen detection by PCR solely enables identification of the pathogen (Schreiber et al., 2013). Numerous studies comparing both methods conclude that the time BCs require to provide positive results is too slow for guiding therapy (Westh et al., 2009; Bloos et al., 2010; Lehmann et al., 2010; Schreiber et al., 2013). Thus, PCR-based assays, which exhibit a turnaround time of several hours may be an important additive tool (Lehmann et al., 2010).

Both methods, BC and PCR, identify the microorganisms directly in the blood. However, at the time of diagnosis, the pathogen may have left the bloodstream, while it still triggers the dysregulated response of the immune system of the host. Thus, another promising approach is to analyze the immunological imprint of the pathogen and infer the pathogen type based on the transcriptional response to the infection. Previous studies have shown that genome-wide transcriptome analysis facilitates the identification of genes with specific expression signatures in sepsis data (Prucha et al., 2004; Shanley et al., 2007). As these genes quantify the state of acute sepsis, they can be considered as biomarkers for this condition. Other research groups used biomarkers to distinguish the microorganisms causing the infection, or to predict the survival chances of infected patients (Pachot et al., 2006; Pankla et al., 2009). Furthermore, septic shock patients have been successfully classified into subgroups using whole-blood gene expression data from microarrays (Wong et al., 2010). Therefore, incorporation of host response transcription data holds great potential to get insights into the systemic host reaction, thus leading to an improved pathogen detection and differentiation. Especially with respect to the rapid increase in incidence of fungal induced sepsis cases, an early detection of fungal sepsis would be of great value.

The genome-wide approach of this study provides an unbiased screening. This strategy facilitates the identification of transcriptional biomarkers featuring distinct expression signatures depending on whether the infectious pathogen is of bacterial or fungal origin. A classifier based on these biomarkers enables the classification of causative microorganisms in new samples. Here, we apply a whole-genome approach for screening the transcriptional response to blood infections and to identify biomarkers. For clinical application, however, a technology like western blot or PCR, which is faster and more accurate or relevant would be advantageous for measuring expression intensities of the biomarker genes. Nevertheless, the present study gives a starting point for the development of a classification device such as a biochip. We based this work on a whole-blood model, as this model takes the *in vivo* complexity of immune responses into account and, compared to other model organisms, the blood components are similar to the human organism with respect to their abundance and functioning (Maccallum, 2012; Hünniger et al., 2014).

2. Materials and Methods

2.1. Microarray Data Generation and Preprocessing

A human whole-blood model was used as described previously (Hünniger et al., 2014). Briefly, HBSS (for mock-infected control) or the human pathogenic fungi *Candida albicans* SC5314 (Gillum et al., 1984) and *Aspergillus fumigatus* ATCC46645 (each 1×10^6/ml), the Gram-positive bacterium *Staphylococcus aureus* ATCC25923 (1×10^6/ml) and the Gram-negative bacterium *Escherichia coli* ATCC25922 (4×10^3/ml) were added to anticoagulated blood of healthy human donors (male, ≤40 years of age) and incubated at 37°C with gentle rotation for 4 or 8 h. The samples of all pathogens cover three or four different donors with one or two samples each. Infected blood was collected and stored in PAXgene Blood RNA Tubes (PreAnalytiX) to stabilize intracellular RNA until further use. RNA isolation was performed using the PAXgene Blood RNA Kit (PreAnalytiX) corresponding to the manufacturer's instruction. The Illumina® TotalPrep™RNA Amplification Kit (Ambion) was used for RNA amplification and cRNA transcription. RNA concentrations and quality were assessed by NanoDrop 1000 (Thermo Scientific) and Agilent 2100 Bioanalyzer (Agilent Technologies). Expression levels of RNA samples were analyzed with Illumina® HumanHT-12 v4 Expression BeadChip Kit (Illumina) following manufacturer's protocol. The chip data was background corrected and log-transformed by applying the functions "lumiR" and "lumiT" of the R package "lumi" (Du et al., 2008). Genes with a detection $p < 0.01$ in at least one sample were considered as expressed. Putative and/or not well-characterized genes (i.e., gene symbols starting with ENSG, NT_, LOC, MGC, HS., FLJ, KIAA, or CxORF) were removed, leaving 10449 genes for analysis. The microarray data have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002), accession number GSE65088 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65088).

2.2. Reference Genes Based Normalization

The normalization followed the approach of Vandesompele et al. (2002) which is based on non-normalized expression values of all samples. From a list of putative control genes covering housekeeping genes and reference genes suggested previously by Stamova et al. (2009) and Kwon et al. (2009), genes with most stable expression were selected. First, the gene stability measure M as introduced by Vandesompele et al. was calculated for each control gene as the average pairwise variation of a gene, i.e., the pairwise standard deviation of the ratios of the control gene to all other control genes. Thus, genes with lower M values are associated with a more stable expression. Iteratively, the gene with the largest M value was removed and the calculation was repeated. In this way, a ranking of genes was obtained, representing their stability. The geometric mean of the expression values of the n best ranked genes was then used as normalization factor (NF_n)—as a vector for all samples.

Initially, the three most stable genes (NF_3) were used to determine the optimal number of genes for NF calculation. Then, more genes were successively included (NF_4, NF_5, ...) as long as the inclusion leads to significant changes on the normalization factor. To quantify these changes, the pairwise variations of each two consecutive NFs were computed. As threshold, 0.15 was used as recommended by Vandesompele et al. A value surpassing this threshold indicate that the inclusion of another gene into calculation is necessary.

2.3. Selection of Differentially Expressed Genes

Differentially expressed genes were determined using the Bioconductor package "limma" (Gentleman et al., 2004; Smyth, 2005) of the statistical programming language R. Limma fits linear models to the expression values of each gene and determines differential expression using moderated t-statistics. P-values were adjusted according to the method of Benjamini and Hochberg (1995). Genes with an adjusted $p < 0.05$ and a $\log_2$-fold change of at least ± 1 were regarded as differentially expressed.

2.4. The Random Forest Classifier

The random forest classifier was built using the "randomForest" package (Liaw and Wiener, 2002) for the R programming language. There are two main parameters which may influence the performance of the classifier: *ntree* and *mtry*. While *ntree* describes the number of trees that are built by the random forest algorithm, *mtry* represents the number of genes used at each split when building a tree. Svetnik et al. (2004) and Díaz-Uriarte and Alvarez de Andrés (2006) showed that the random forest algorithm features high predictive performance, even without parameter adjustment. Only the number of trees needs to be sufficiently large to get stable results. Therefore, the random forest classifier was built growing 100,000 trees. A cross-validation examining the effect of changing *mtry* and *ntree* showed that altering the parameters has no effect on the classification accuracy (Supplementary Material). Thus, we kept the parameter *mtry* on its default value, which is $\lfloor \sqrt{g} \rfloor$, where g is the number of genes of the input dataset.

For the selection of biomarker genes, the measure "mean decrease in accuracy" was used for determining the variable importance values for each gene. The importance values were computed for each class (fungal, bacterial, and mock-infected class) by building random forests with 100,000 trees. The normalized dataset, which was reduced to the data of differentially expressed genes, was used as input.

We scaled the certainty score to a range from 0 to 1. Before scaling, the score represents the proportion of class predictions from all trees of the random forest, which yield the same class as the final classification by the classifier. Let p be this proportion and let N be the number of possible classes (in this study, $N = 3$, as we consider a fungal, a bacterial, and a mock-infected class), then the certainty score is calculated as

$$certainty\ score = \frac{p - \frac{1}{N}}{1 - \frac{1}{N}}. \quad (1)$$

2.5. Performance Assessment

The *C. neoformans* (strain H99, provided by Robin May, University of Birmingham) dataset was generated identical to the other fungi data and quantified using the same chip technology. Expression levels were measured 4 h (3 donors) and 8 h (3 donors) post infection. Mock-infected control samples were simultaneously produced. Before classification, the expression intensities were normalized based on the reference genes which were determined previously without the *C. neoformans* data (**Figure 1**).

Multidimensional scaling (MDS) was performed using the "cmdscale" function of R. After determining the Spearman correlation of the samples of the normalized *C. neoformans* dataset, the Euclidean distances between these samples were calculated based on the correlation matrix and used as input for the MDS computation. In this way, samples with high correlations are close to each other in the MDS plot.

3. Results

3.1. Reference Genes Based Normalization

Our first step in building a classifier which discriminates between bacterial and fungal infection is to normalize gene expression values with help of reference genes (**Figure 1**). The motivation of using reference genes instead of the control samples for normalization is that our classifier should be able to be applied in clinical settings, i.e., for patients, where no control samples exist. To identify reference genes, we used a knowledge driven and data driven approach. First, we considered 10 known housekeeping genes as well as 17 reference genes which were previously suggested by Kwon et al. (2009) and Stamova et al. (2009) (**Table 1**). Next, we checked which of those genes have a stable expression profile within our dataset. Therefore, we followed the method proposed by Vandesompele et al. (2002), where the stability of a gene is determined on the basis of ratios of raw gene expression values (Materials and Methods). The normalization factor (NF) is then calculated as the geometric mean of the most stably expressed reference genes.

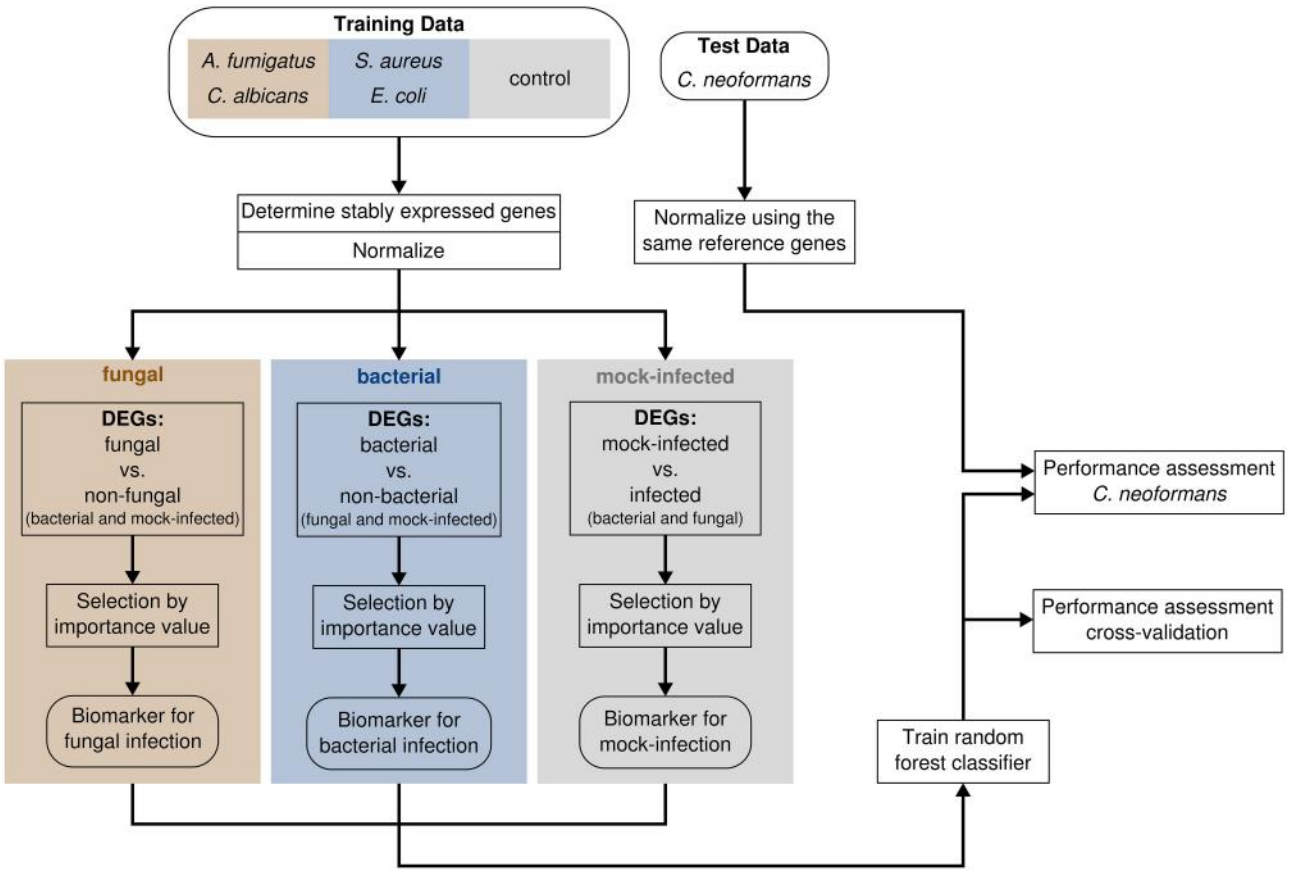


FIGURE 1 | The workflow for biomarker identification, classifier construction and performance assessment.

From the 27 considered genes, we determined *CTBP1*, *TBP*, and *CRY2* as the most stable ones. When comparing the pairwise variations of all successive NFs, we found that using only the three most stably expressed genes is sufficient for producing an accurate NF (Supplementary Figure 2). Including a fourth reference gene leads to no significant changes of the NF, indicated by a low pairwise variation of 0.0496. This value is below the threshold of 0.15 that was recommended by Vandesompele et al. for including more genes. Furthermore, the Spearman correlation between NF_3 and NF_4 is >0.99, which also demonstrates that considering a fourth gene is not necessary.

3.2. Selection of Biomarker Genes

The identification of biomarkers, i.e., genes with a specific expression pattern in case of a whole-blood infection, requires the reduction of the gene set by so called feature selection. As gene expression data is high-dimensional by nature, feature selection is one of the most important tasks when building a classifier based on genome wide transcription data. The aim of feature selection is to pick the most informative genes and to remove irrelevant predictors, thus resulting in a dimension reduction. In this way, we can reduce the complexity of the classification while at the same time the predictive performance can be increased. In general, we can distinguish three types of feature selection: filter methods, wrapper methods, and embedded methods (Saeys et al., 2007).

We performed feature selection using the filter and the embedded approach by first determining differentially expressed genes (DEGs) and then selecting genes which are most important for accurate classification (**Figure 1**). To identify genes showing different expression patterns between the pathogen types rather than between the species, we grouped data into three classes. The fungal species *C. albicans* and *A. fumigatus* form the class "fungal," while the bacterial species *S. aureus* and *E. coli* were assembled to the "bacterial" class. The samples of the control group are represented by the class "mock-infected."

3.2.1. Selection of Differentially Expressed Genes

To identify transcriptional responses related to blood infection by fungi or bacteria we determined DEGs for the three classes. A gene is regarded as a DEG for one class, if its expression levels are significantly different to both other classes merged together (Materials and Methods). In this way, we found 204 DEGs for the fungal class, 184 for the bacterial class, and 150 for the mock-infected class. Of these genes, 68 were identified as differentially expressed in all 3 classes simultaneously. The union of the three sets of DEGs comprises a total of 402 genes.

TABLE 1 | Housekeeping genes and putative reference genes suggested by other studies were used as input for determining stably expressed reference genes.

Housekeeping genes listed at Vandesompele et al.	Reference genes suggested by Stamova et al.	Reference genes suggested by Kwon et al.
ACTB	*TRAP1*	*ZNF207*
B2M	*DECR1*	*OAZ1*
GAPDH	*FPGS*	*LUC7L2*
HMBS	*FARP1*	***CTBP1***
HPRT1	*MAPRE2*	*TRIM27*
RPL13A	*PEX16*	*GPBP1*
SDHA	*GINS2*	*ARL8B*
TBP	***CRY2***	*UBQLN1*
UBC	*CSNK1G2*	*PAPOLA*
YWHAZ	*A4GALT*	*CUL1*
		DIMT1L
		FBXW2
		SPG21

The symbols in the genes FPGS, FARP1, PEX16, GINS2, A4GALT, and SPG21 could not be found in our dataset and thus were not considered. The genes exhibiting the most stable expression are bolded.

3.2.2. Selection by Importance Value

We further reduced the set of DEGs to genes being most important for accurate classification. To identify these genes, we used the variable importance measure integrated in the random forest algorithm (Materials and Methods). We selected the top 11, 6, and 21 genes for the classes fungal, bacterial, and mock-infected, respectively, as these genes form groups covering the highest importance values (**Figure 2**). They are biomarkers for their respective group of pathogens.

3.2.3. Functional Annotation of Selected Biomarker Genes

To get insights into the function of the biomarker genes, we performed a Gene Ontology (GO) (Ashburner et al., 2000) enrichment analysis. We employed the tool "GOrilla" (Eden et al., 2009) to identify over-represented GO categories. This web-based tool uses an hypergeometric model to test for enrichment and performs *p*-value adjustment for multiple testing according to the false discovery rate.

At a significance level of 0.05 we found 32 enriched GO terms connected to the identified biomarker genes (Supplementary Table 2). The list comprises terms from the areas of signal transduction, activation of the immune system, response to cytokine stimuli, and down-regulation of phosphorylation. Besides that, GOrilla also identified the category "regulation of sequence-specific DNA binding transcription factor activity" as over-represented. Although numerous of the enriched GO terms are connected to the immune response, we found that multiple biomarkers are related to other processes. For example, genes are involved in cellular growth (*TBC1D7*, *GADD45B*), vesicle transport (*VPS18*), cell proliferation (*PIM1*, *PIM3*), cell adhesion (*VCAN*), ion transport (*FXYD6*), or iron uptake (*TFRC*).

Many genes of our biomarkers are already linked to sepsis by other studies. While *IL6* was previously identified as biomarker for sepsis (Pierrakos and Vincent, 2010), *GADD45B*, *SOCS3*, and *IRG1* were shown to be up-regulated in septic patients (Johnson et al., 2007; Li et al., 2013). Moreover, it has been shown that *IL1F9* is up-regulated by *S. aureus* cell wall proteins in human peripheral blood mononuclear cells (Kang et al., 2012). Furthermore, *RGS1*, *CCL3*, and *SOCS1* were connected to sepsis in animal studies (Panetta et al., 1999; Takahashi et al., 2002; Grutkoski et al., 2003), while for *CTSD* increased expression levels were observed in mice with induced septic shock (Yoo et al., 2013). *MAP3K8* is linked to sepsis in mice, with being crucial for the TNF production (Mielke et al., 2009). Furthermore, the gene *MIR155HG* showed significantly higher expression values in samples with bacterial or fungal infection than in the mock-infected controls. This gene encodes for the microRNA miR-155, which is known to be involved in the regulation of antimicrobial immune response (O'Connell et al., 2007; Rodriguez et al., 2007; Das Gupta et al., 2014).

Examining the expression signatures of the selected genes (**Figure 3**, Supplementary Figure 1), we discovered that for the fungal and bacterial class, most genes are up-regulated, compared to the respective other two classes. Of the six biomarkers for bacterial blood infection, only one gene (*CXXC5*) was down-regulated, while the other five genes showed up-regulation. For the fungal class, all 11 selected genes were up-regulated. We observed different patterns for the genes of the mock-infected class. Twenty of the 21 genes were down-regulated in the control samples and one gene (*VCAN*) was up-regulated.

Taken together, our feature selection approach was able to identify biomarker genes, which have been shown to be involved in sepsis and also cover a broad range of biological processes.

3.3. Building the Classifier

To determine if an infecting pathogen of an unknown whole-blood sample is of fungal or bacterial origin, the sample is classified using the expression data of the selected biomarkers. We accomplish the classification by a random forest (Breiman, 2001) classifier (the classifier can be found as R object as supplementary file). Random forest is based on an ensemble of decision trees, where each tree is built on a different random subset of the input data. The output of the classifier is determined by the majority vote of the class predictions of all trees. As we used 100,000 trees, the algorithm provides us with 100,000 single classifications. We utilized the votes of the trees to introduce a certainty score for the final classification. This score represents the fraction of class predictions identical with the final classification and was scaled to a range from 0 to 1 (Materials and Methods). In case of a certainty score of 1, all trees have predicted the same class for a given sample and consequently this class was then output by the classifier. On the other hand, the certainty score is 0, if all tree votes are equally distributed across all possible classes. Thus, the score indicates, how sure the classifier is about its decision. Calculating the certainty score for the training data, we achieved average values of 0.941, 0.966, and 0.99 for fungal, bacterial, and mock-infected class, respectively.

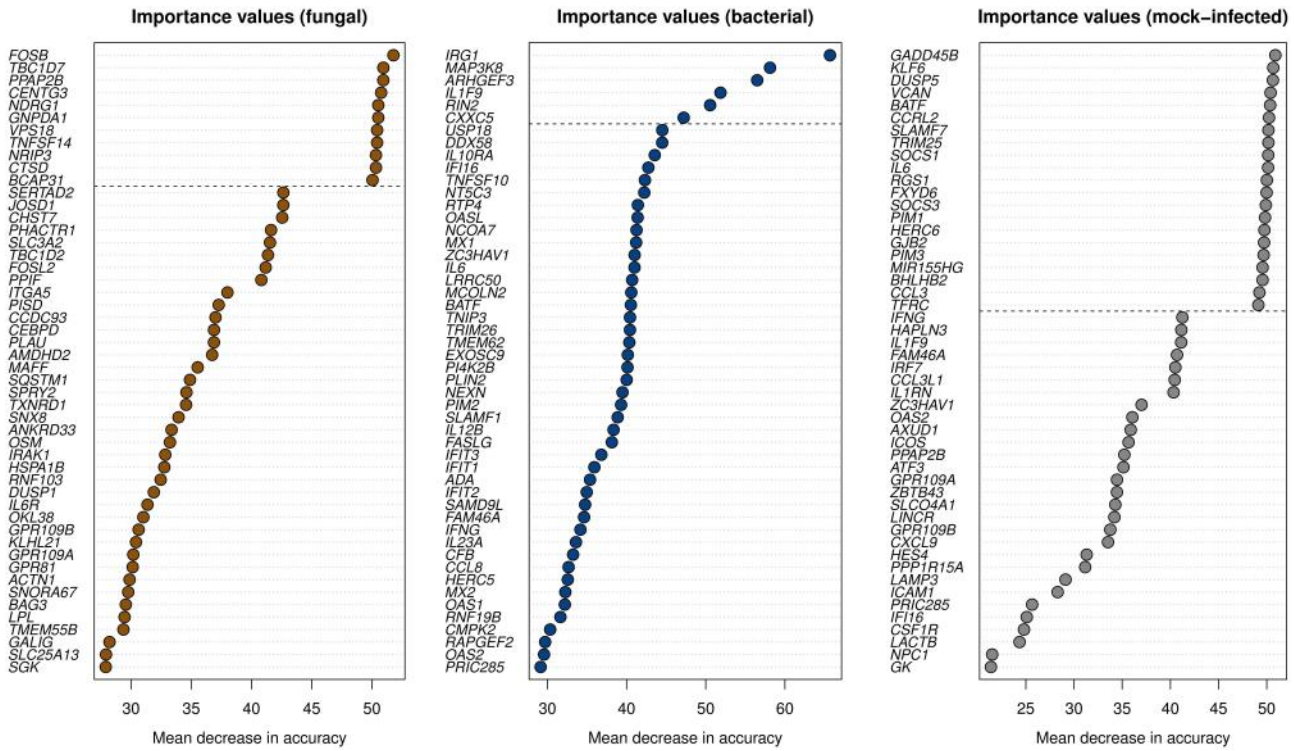


FIGURE 2 | The variable importance values were computed by the random forest algorithm. A gene with larger values exhibits a higher influence on the correct class predictions. The 50 highest importance values of the measure "mean decrease in accuracy" are shown. Genes above the dashed lines were selected as biomarkers for the corresponding classes.

3.4. Performance Assessment

Having built our classifier, we next studied its performance in distinguishing between fungal or bacterial blood infection. Our aim was to accurately classify new samples by the given classification model. Therefore, the performance assessment methods have to yield unbiased accuracy rates. To get unbiased estimates of accuracy, the samples for testing the classifier should be independent from the samples for training the classifier. We fulfilled this requirement with additionally independently created data comprising RNA expression measurements of human whole-blood samples infected with *C. neoformans*. An additional approach to assess a classifiers performance is cross-validation. Cross-validation emulates independent test sets in an iterative technique and in this way resolves the need for true test data. Furthermore, we evaluate the ability of the classifier to handle fluctuations in the expression values by classifying samples after adding random noise to the data (Supplementary Material).

3.4.1. Test Data of *C. neoformans*

To assess the performance of the classifier on an independent test set, we created a new dataset of RNA expression measurements of human whole-blood infected with *C. neoformans*. The data comprises 6 samples of fungal infection and 6 mock-infected controls (Materials and Methods). Being part of the phylum of Basidiomycota, *C. neoformans* is a phylogenetically and morphologically very different fungus compared to *C. albicans* and *A. fumigatus*, both belonging to the phylum of Ascomycota (James et al., 2006).

When assessing the classification performance using the new data, our model correctly classified 5 of the 6 fungal samples (83.3%). One sample was wrongly classified as mock-infected. All classifications of the mock-infected samples were performed correctly. In this way, we achieved an overall accuracy rate of 91.7%. The sensitivities are 83.3 and 100%, while the specificities are 100 and 83.3% for fungal and mock-infected class, respectively (**Table 2**). We examined the misclassification in more detail by a correlation analysis using a multidimensional scaling (MDS) plot (**Figure 4**). MDS is a dimension reduction technique, producing an easy-to-visualize output showing relationships within the data. The plot revealed that the misclassified sample shows more similarity to the data of mock-infected class than the other *C. neoformans* samples.

The difference in the accuracy values between the two classes is also reflected in the certainty scores. We obtained an average certainty of 0.475 ($\pm$ 0.190) for all fungal samples, whereas for the mock-infected samples we achieved an average score of 0.810 ($\pm$ 0.165). When splitting the fungal specimen into falsely and correctly classified ones, the observed certainty value for the misclassified sample is higher, 0.654, than for the right classifications, 0.439.

3.4.2. Cross-Validation

When the sample size of a study is relatively small, it is preferred to use all available samples in feature selection and training. However, this leads to a lack of test data. Cross-validation

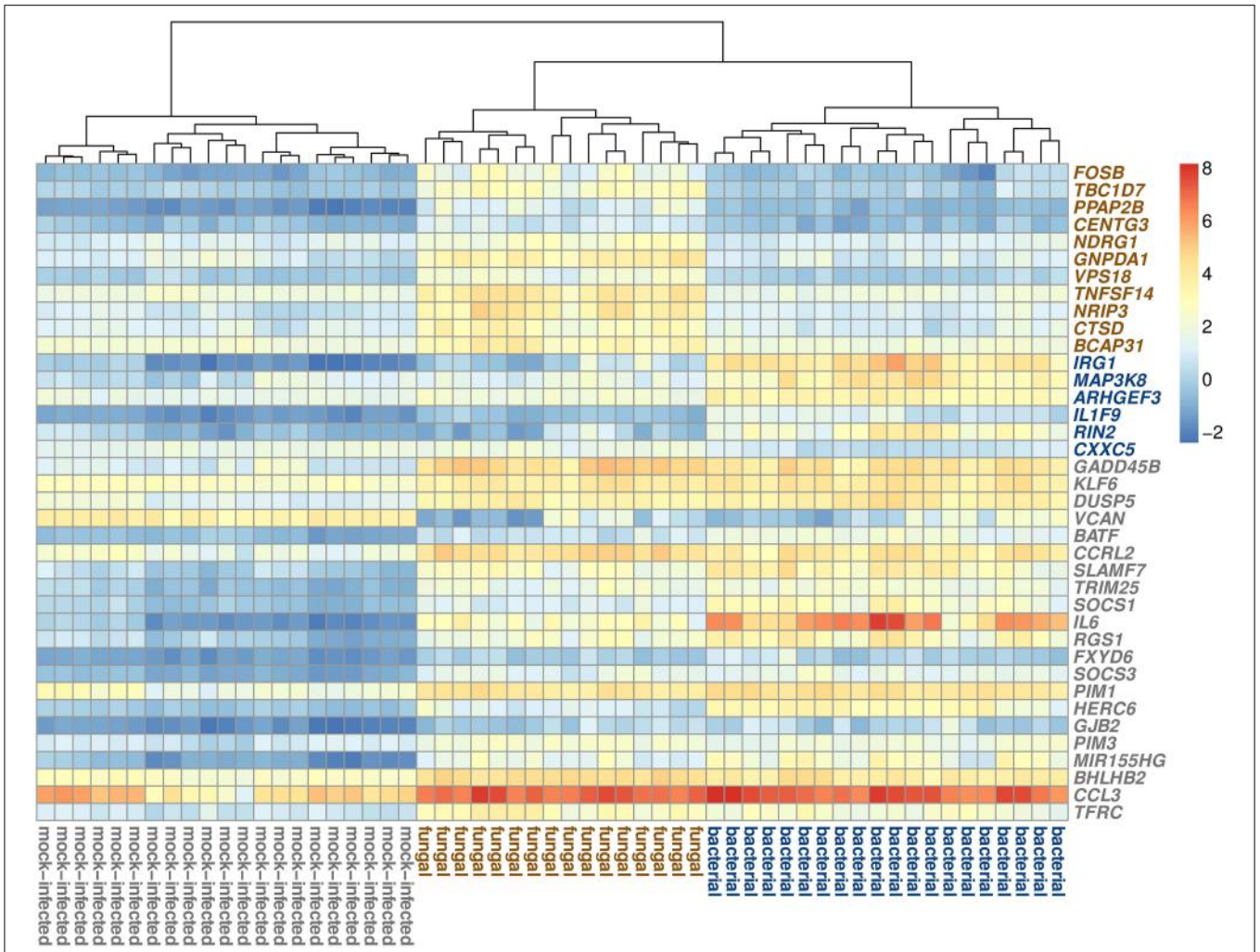


FIGURE 3 | Visualization of the expression patterns of the biomarker genes. The samples are clustered according to their corresponding classes. The heatmap colors correlate with the normalized expression intensities (see key on right side). The colors of the gene symbols indicate the class for which the gene was selected as biomarker (brown = fungal class, blue = bacterial class, gray = mock-infected class).

TABLE 2 | Sensitivities and specificities for the performance assessments.

	Sensitivity			Specificity		
	Bacterial	Fungal	Mock-infected	Bacterial	Fungal	Mock-infected
C. neoformans predictions	–	0.833	1.000	–	1.000	0.833
Cross-validation	0.950	0.938	1.000	0.973	0.976	1.000

The C. neoformans dataset does not comprise samples of the bacterial class. Thus, no sensitivity and specificity could be calculated for this condition.

is a widely used method to overcome this problem by emulating independent test sets without using additional datasets. It works by iteratively setting aside samples for testing, while the remaining samples are used to train the model. The split is performed in the way that each sample of the data is exactly once in the test set. In this way, cross-validation guards against overfitting.

To estimate how accurate the classifier will perform on independent data, we carried out a stratified 10-fold cross-validation (CV). It is important that CV encompasses all feature selection steps, as otherwise a selection bias is induced (Ambroise and McLachlan, 2002). Therefore, we conducted the following procedures on the training set in each CV iteration: determine DEGs, rank the DEGs according to their importance value, select the top-scoring genes, and train a random forest classifier.

In compliance with the CV procedure, the class of each sample of our dataset was predicted and the accuracy of the classification model was estimated. Of the 57 samples, only two were misclassified, while 55 classifications were correct. The two wrong classifications appeared for one bacterial and one fungal sample. All data of the mock-infected class was classified correctly. Thus, the average accuracy of the CV is 96.49% (sensitivities: 93.8, 95, 100% for fungal, bacterial, and mock-infected class; specificities: 97.6, 97.3, 100% for fungal, bacterial, and mock-infected class;

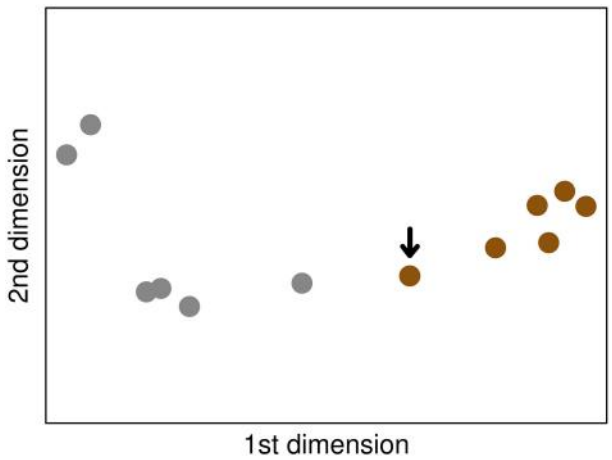


FIGURE 4 | The MDS plot based on the *C. neoformans* dataset, where the relative positions in the plot represent the Euclidean distances of the Spearman correlations of the samples. Small distances correspond to high correlation coefficiens. Brown and gray circles indicate samples of the fungal and the mock-infected class, respectively. The arrow marks the fungal sample that was misclassified as mock-infected control.

Table 2). The average certainties of the classifications were 0.795 ($\pm$ 0.169), 0.855 ($\pm$ 0.18), and 0.937 ($\pm$ 0.085) for the classes fungal, bacterial, and mock-infected, respectively.

4. Discussion

Here we present an transcriptome analysis of human whole-blood data comparing bacterial and fungal infections with mock-infected control samples. Based on the regulatory differences, we identified biomarker genes, which show characteristic expression patterns according to their respective causative pathogen type. The selection was not only based on statistical significance. It also took into account to what extent the random forest classification algorithm assesses these genes as important for separating the given classes. In this way, we applied two different methods of feature selection: the filter approach and the embedded approach. With the detection of differentially expressed genes we are able to remove most of the irrelevant genes and extract a set of potential transcriptional marker genes. The selection by differential expression is a widely used method for identifying sepsis related marker genes (Prucha et al., 2004; Pachot et al., 2006; Shanley et al., 2007; Pankla et al., 2009). The subsequent calculation of gene importance values using the random forest algorithm allows us to identify the genes showing the strongest and most constant up- or down-regulation as a consequence of the blood infection by the particular type of microorganisms. In this way, we were able to remarkably reduce a set of whole-genome expression measurements to significant signatures distinguishing bacterial from fungal infections and mock-infected controls. The genes identified as biomarkers for the mock-infected class exhibit similar signatures for both infection types, fungal and bacterial. Most of these genes show down-regulation in the mock-infected samples. However, at the same time they were up-regulated in the infected samples, irrespective of the infecting pathogen type. Therefore, they possibly reflect cellular regulations to respond microbial infections in general. Thus, they can be considered as pathogen-independent markers for whole-blood infections. Studies investigating a broader range of pathogens should be carried out to confirm this hypothesis.

Using a human whole-blood model in this work is supported by several advantages. First, as opposed to purified human immune cells, it also considers the *in vivo* complexity of the immune response in blood (Hünniger et al., 2014). Next, there are no differences in proportions and functioning of the peripheral blood components between this model and the target organism, the human, in contrast to other model organisms like mice (Maccallum, 2012). Furthermore, human whole-blood infection models have been successfully used previously to identify factors of virulence (Echenique-Rivera et al., 2011) and to analyze human immune responses (Tena et al., 2003).

Following a genome-wide approach allows us to consider all genes as potential biomarkers for pathogen type recognition, even if they are not related to immune response. Therefore, with respect to the screening for biomarkers, using a whole-genome method is more promising than techniques which are limited to a small number of candidates, like serum cytokine analysis. Indeed, the selected biomarker genes cover a broad range of functions. In this way, these genes may facilitate the recognition of bloodstream infections even when the immune system of the patient is affected by additional diseases. Besides that, we found the gene *MIR155HG* as up-regulated in the samples with infections. Recently, Das Gupta et al. (2014) have shown that miR-155 up-regulation is not specific to host response on bacterial pathogens. They also detected increased expression levels as reaction to *A. fumigatus* infections. As we observed up-regulations for all considered species, fungi as well as bacteria, our results confirm the findings that miR-155 is involved in a general host response to infections, covering a wide range of pathogens. Besides, numerous of the selected biomarkers were previously associated to sepsis in either human or animal studies. This finding indicates, that although our results are based on an experimental model instead of patient data, we could identify characteristic gene regulations in response to microbial bloodstream infections.

Preceding the feature selection steps, we successfully identified the three most stable genes from a set of published control genes and used them as reference for normalizing the dataset. In this way, we do not use absolute gene expression values to train our classifier. Instead, we use expression values relative to the geometric mean of the reference genes. Regarding the application case, a user of the classifier aims to identify the pathogen type using only a single blood sample without mock-infected controls for comparison. It is well known that the intensity values on microarrays are influenced by technical variations and errors connected with wet lab hand handling of samples as well as hybridization and scanning of the chip. These differences can not be detected on a single sample, but they do affect the absolute intensity values. With normalizing relative to reference genes, we control for this effect, as all genes on the chip are influenced in the same way. Furthermore, this method can easily be adapted to other quantification methods like PCR.

Using the biomarker genes, we trained a random forest classifier to classify the pathogen type in whole-blood samples. Random forest provides several advantages making it suitable for this study. It is fast in training and testing, supports multiclass classifications and provides the variable importance for evaluating the input features. With this embedded measure, we were able to select the best class-separating genes leading to a small set of biomarkers. There are further classification methods like support vector machines or naïve Bayes classifiers, which were successfully applied on microarray data in other studies (Kelemen et al., 2003; Howrylak et al., 2009). For comparison, we tested the classification performance of these two techniques on both the *C. neoformans* dataset and the cross-validation, using the previously selected biomarkers (Supplementary Material). The support vector machine as well as the naïve Bayes method yielded the same classifications of all samples as the random forest model. The fact that the three classification methods are very different in their functional principles and the results are unaffected by the choice of the model indicates that the selected biomarker genes are robust.

The certainty score based on the votes of the trees provides an easy-to-compare measure for assessing the classification quality. It directly reflects the ability of the classification model to properly classify the input data. This means, a class prediction with a high certainty score is more likely to be correct, than one with a low score. One possible application case for this measure is the introduction of a threshold, followed by the removal of low-scoring classifications.

We tested the classifier with an additional dataset comprising whole-blood samples of fungal infection and mock-infected controls. The medically important fungus used for these additional samples, *C. neoformans*, is phylogenetically very different from *C. albicans* and *A. fumigatus*. These differences can lead to varieties in the transcriptional response of the host. However, the accuracy value of about 92% indicate that the selected biomarker genes are largely unaffected. Therefore, these genes are general indicators for whole-blood infections caused by fungi. The MDS analysis revealed that the misclassified fungal sample shows a greater similarity to the specimen of the mock-infected class than to the fungal cases. Although the divergence with the other fungal samples is only small, the differences are sufficient for wrong classification. Consequently, the correct classifications of the *C. neoformans* samples are possibly unsure. Indeed, the certainty values are much lower for the fungal class, compared to the mock-infected controls. Furthermore, we were surprised to find the certainty score of the misclassified sample being higher than the average score of the remaining fungal specimen. This observation confirms the assumption that the prediction of *C. neoformans* as fungal infected blood sample is a difficult task for the classifier, but still leads to mostly correct results.

High accuracy values were not only achieved when validating the classifier with the additional *C. neoformans* dataset, but also when testing it with stratified 10-fold CV. This broadly used performance assessment technique iteratively estimates the accuracy of a prediction model without an independent dataset. The two misclassifications in this test appeared for fungal and bacterial class. The predictions of the fungal and the bacterial class also exhibit the lowest values and the largest fluctuations of the certainty scores. However, it should be noted that the average scores are still high, as 0.795 is the smallest of them.

In summary, the results of the assessments by using an additional dataset of fungal infection, i.e., the external validation, as well as by performing a CV, i.e., the internal validation, are promising. Most of the tested samples were correctly classified, although in some cases right classifications were accompanied by low certainty scores.

We also performed a noise-robustness test to examine whether the classifier can compansate fluctuations in the expression data. The high accuracy rates indicate that the indentified biomarkers are robust with respect to changes in their expression intensities. This robustness is important for a potential clinical application, where patients are of different age, sex, medication, and health condition and thus expression intensities of the same genes will vary between these patients.

The experimental model of this work comprises the infection of blood from healthy human donors with typical sepsis causing microorganisms. Although we gained important insights into the transcriptional response on the pathogens, our findings possibly can not be directly utilized for clinical application. To achieve that, further analyses on gene expression data from septic patients as well as functional follow-up studies have to be performed. Unfortunately, whole-genome expression data from septic patients where the causing pathogen is known is rare in publicly accessible databases. Especially, datasets comprising the transcriptional response to fungal induced sepsis are scarce. Thus, we lack the basis for more clinical relevant investigations, which is why it remains an open task for future research. Furthermore, it should be noted that the presented classifier can not be used to identify the infecting species. Rather it is supposed to answer the question if the pathogen is of bacterial or fungal origin and whether or not it is necessary to administer antimycotics instead of antibiotics. To initiate a species dependent therapy, more requirements have to be fulfilled, e.g., in case of a bacterial infection, the appropriate antibiotic has to be determined by an antibiogram.

In this study we present an effective selection of genes showing characteristic expression patterns depending on the type of the infectious organism. The resulting small gene set was used to train a fast and accurate random forest classifier, which performs well in predicting the class of the pathogen. Examining the transcriptional footprint of the sepsis causing microorganism in the blood of the host is a promising approach for quick pathogen identification. With the presented classification model we meet the increasing challenge of fungal induced septic infections requiring novel detection methods.

Author Contributions

AD did the bioinformatic analysis and co-wrote the manuscript. KH performed the experiments, generated the data, and co-wrote the manuscript. MW discussed the analysis and co-wrote the

manuscript. RG, OK, and JL designed the research and co-wrote the manuscript.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Centre/Transregio 124 FungiNet (subprojects B3, INF, C3) as well as German Ministry for Education and Science in the program Unternehmen Region (BMBF 03Z2JN21).

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00171/abstract

References

Ambroise, C., and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6562–6566. doi: 10.1073/pnas.102102699

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soci. Ser. B* 57, 289–300.

Bloos, F., Hinder, F., Becker, K., Sachse, S., Mekontso Dessap, A., Straube, E., et al. (2010). A multicenter trial to compare blood culture with polymerase chain reaction in severe human sepsis. *Intensive Care Med.* 36, 241–247. doi: 10.1007/s00134-009-1705-z

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Das Gupta, M., Fliesser, M., Springer, J., Breitschopf, T., Schlossnagel, H., Schmitt, A.-L., et al. (2014). *Aspergillus fumigatus* induces microRNA-132 in human monocytes and dendritic cells. *Int. J. Med. Microbiol.* 4, 2–6. doi: 10.1016/j.ijmm.2014.04.005

Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinform.* 7:3. doi: 10.1186/1471-2105-7-3

Du, P., Kibbe, W. A., and Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547–1548. doi: 10.1093/bioinformatics/btn224

Echenique-Rivera, H., Muzzi, A., Del Tordello, E., Seib, K. L., Francois, P., Rappuoli, R., et al. (2011). Transcriptome analysis of Neisseria meningitidis in human whole blood and mutagenesis studies identify virulence factors involved in blood survival. *PLoS Pathog.* 7:e1002027. doi: 10.1371/journal.ppat.1002027

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinform.* 10:48. doi: 10.1186/1471-2105-10-48

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Engel, C., Brunkhorst, F. M., Bone, H.-G., Brunkhorst, R., Gerlach, H., Grond, S., et al. (2007). Epidemiology of sepsis in Germany: results from a national prospective multicenter study. *Intensive Care Med.* 33, 606–618. doi: 10.1007/s00134-006-0517-7

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Gillum, A. M., Tsay, E. Y. H., and Kirsch, D. R. (1984). Isolation of the *Candida albicans* gene for orotidine-5'-phosphate decarboxylase by complementation of S. cerevisiae ura3 and E. coli pyrF mutations. *Mol. Gen. Genet.* 198, 179–182. doi: 10.1007/BF00328721

Grutkoski, P. S., Chen, Y., Chung, C. S., and Ayala, A. (2003). Sepsis-induced SOCS-3 expressio n is immunologically restricted to phagocytes. *J. Leukoc. Biol.* 74, 916–922. doi: 10.1189/jlb.0303108

Howrylak, J. A., Dolinay, T., Lucht, L., Wang, Z., Christiani, D. C., Sethi, J. M., et al. (2009). Discovery of the gene signature for acute lung injury in patients with sepsis. *Physiol. Genomics* 37, 133–139. doi: 10.1152/physiolgenomics.90275.2008

Hünniger, K., Lehnert, T., Bieber, K., Martin, R., Figge, M. T., and Kurzai, O. (2014). A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. *PLoS Comput. Biol.* 10:e1003479. doi: 10.1371/journal.pcbi.1003479

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., et al. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* 443, 818–822. doi: 10.1038/nature05110

Johnson, S. B., Lissauer, M., Bochicchio, G. V., Moore, R., Cross, A. S., and Scalea, T. M. (2007). Gene expression profiles differentiate between sterile SIRS and early sepsis. *Ann. Surg.* 245, 611–621. doi: 10.1097/01.sla.0000251619.10648.32

Kang, S.-S., Kim, H. J., Jang, M. S., Moon, S., In Lee, S., Jeon, J. H., et al. (2012). Gene expression profile of human peripheral blood mononuclear cells induced by *Staphylococcus aureus* lipoteichoic acid. *Int. Immunopharmacol.* 13, 454–460. doi: 10.1016/j.intimp.2012.05.010

Kelemen, A., Zhou, H., Lawhead, P., and Liang, Y. (2003). "Naive Bayesian classifier for microarray data," in *Proceedings of the International Joint Conference on Neural Networks, 2003* (Portland, OR), 1769–1773.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., et al. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit. Care Med.* 34, 1589–1596. doi: 10.1097/01.CCM.0000217961.75225.E9

Kwon, M. J., Oh, E., Lee, S., Roh, M. R., Kim, S. E., Lee, Y., et al. (2009). Identification of novel reference genes using multiplatform expression data and their validation for quantitative gene expression analysis. *PloS ONE* 4:e6162. doi: 10.1371/journal.pone.0006162

Lehmann, L. E., Hunfeld, K.-P., Steinbrucker, M., Brade, V., Book, M., Seifert, H., et al. (2010). Improved detection of blood stream pathogens by real-time PCR in severe sepsis. *Intensive Care Med.* 36, 49–56. doi: 10.1007/s00134-009-1608-z

Li, Y., Zhang, P., Wang, C., Han, C., Meng, J., Liu, X., et al. (2013). Immune responsive gene 1 (IRG1) promotes endotoxin tolerance by increasing A20 expression in macrophages through reactive oxygen species. *J. Biol. Chem.* 288, 16225–16234. doi: 10.1074/jbc.M113.454538

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R. News* 2, 18–22.

Maccallum, D. M. (2012). Hosting infection: experimental models to assay Candida virulence. *Int. J. Microbiol.* 2012:363764. doi: 10.1155/2012/363764

Martin, G. S. (2012). Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. *Expert Rev. Anti Infect. Ther.* 10, 701–706. doi: 10.1586/eri.12.50

Martin, G. S., Mannino, D. M., Eaton, S., and Moss, M. (2003). The epidemiology of sepsis in the United States from 1979 through 2000. *N. Engl. J. Medi.* 348, 1546–1554. doi: 10.1056/NEJMoa022139

Mielke, L. A., Elkins, K. L., Wei, L., Starr, R., Tsichlis, P. N., O'Shea, J. J., et al. (2009). Tumor progression locus 2 (Map3k8) is critical for host defense against Listeria monocytogenes and IL-1 beta production. *J. Immunol.* 183, 7984–7993. doi: 10.4049/jimmunol.0901336

O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G., and Baltimore, D. (2007). MicroRNA-155 is induced during the macrophage inflammatory response. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1604–1609. doi: 10.1073/pnas.0610731104

Pachot, A., Lepape, A., Vey, S., Bienvenu, J., Mougin, B., and Monneret, G. (2006). Systemic transcriptional analysis in survivor and non-survivor septic shock patients: a preliminary study. *Immunol. Lett.* 106, 63–71. doi: 10.1016/j.imlet.2006.04.010

Panetta, R., Guo, Y., Magder, S., and Greenwood, M. T. (1999). Regulators of G-protein signaling (RGS) 1 and 16 are induced in response to bacterial lipopolysaccharide and stimulate c-fos promoter expression. *Biochem. Biophys. Res. Commun.* 259, 550–556. doi: 10.1006/bbrc.1999.0817

Pankla, R., Buddhisa, S., Berry, M., Blankenship, D. M., Bancroft, G. J., Banchereau, J., et al. (2009). Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis. *Genome Biol.* 10:R127. doi: 10.1186/gb-2009-10-11-r127

Pierrakos, C., and Vincent, J.-L. (2010). Sepsis biomarkers: a review. *Crit. Care* 14:R15. doi: 10.1186/cc8872

Prucha, M., Ruryk, A., Boriss, H., Möller, E., Zazula, R., Herold, I., et al. (2004). Expression profiling: toward an application in sepsis diagnostics. *Shock* 22, 29–33. doi: 10.1097/01.shk.0000129199.30965.02

Rittirsch, D., Flierl, M. A., and Ward, P. A. (2008). Harmful molecular mechanisms in sepsis. *Nat. Rev. Immunol.* 8, 776–787. doi: 10.1038/nri2402

Rodriguez, A., Vigorito, E., Clare, S., Warren, M. V., Couttet, P., Soond, D. R., et al. (2007). Requirement of bic/microRNA-155 for normal immune function. *Science* 316, 608–611. doi: 10.1126/science.1139253

Saeys, Y., Inza, I. N., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Schreiber, J., Nierhaus, A., Braune, S. A., de Heer, G., and Kluge, S. (2013). Comparison of three different commercial PCR assays for the detection of pathogens in critically ill sepsis patients. *Med. Klin. Intensivmed. Notfallmed.* 108, 311–318. doi: 10.1007/s00063-013-0227-1

Shanley, T. P., Cvijanovich, N., Lin, R., Allen, G. L., Thomas, N. J., Doctor, A., et al. (2007). Genome-level longitudinal expression of signaling pathways and gene networks in pediatric septic shock. *Mol. Med.* 13, 495–508. doi: 10.2119/2007-00065.Shanley

Smyth, G. (2005). "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0/23

Stamova, B. S., Apperson, M., Walker, W. L., Tian, Y., Xu, H., Adamczy, P., et al. (2009). Identification and validation of suitable endogenous reference genes for gene expression studies in human peripheral blood. *BMC Med. Genomics* 2:49. doi: 10.1186/1755-8794-2-49

Svetnik, V., Liaw, A., Tong, C., and Wang, T. (2004). "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Multiple Classifier Systems*, eds F. Roli, J. Kittler, and T. Windeatt (Berlin; Heidelberg: Springer), 334–343.

Takahashi, H., Tashiro, T., Miyazaki, M., Kobayashi, M., Pollard, R. B., and Suzuki, F. (2002). An essential role of macrophage inflammatory protein 1α/CCL3 on the expression of host's innate immunities against infectious complications. *J. Leukoc. Biol.* 72, 1190–1197. doi: 10.4049/jimmunol.169.8.4460

Tena, G. N., Young, D. B., Eley, B., Henderson, H., Nicol, M. P., Levin, M., et al. (2003). Failure to control growth of mycobacteria in blood from children infected with human immunodeficiency virus and its relationship to T cell function. *J. Infect. Dis.* 187, 1544–1551. doi: 10.1086/374799

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3:RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034

Westh, H., Lisby, G., Breysse, F., Böddinghaus, B., Chomarat, M., Gant, V., et al. (2009). Multiplex real-time PCR and blood culture for identification of bloodstream pathogens in patients with suspected sepsis. *Clin. Microbiol. Infect.* 15, 544–551. doi: 10.1111/j.1469-0691.2009.02736.x

Wong, H. R., Wheeler, D. S., Tegtmeyer, K., Poynter, S. E., Kaplan, J. M., Chima, R. S., et al. (2010). Toward a clinically feasible gene expression-based subclassification strategy for septic shock: proof of concept. *Crit. Care Med.* 38, 1955–1961. doi: 10.1097/CCM.0b013e3181eb924f

Yoo, H., Ahn, E.-R., Kim, S.-J., Lee, S.-H., Oh, S. H., and Kim, S.-Y. (2013). Divergent results induced by different types of septic shock in transglutaminase 2 knockout mice. *Amino Acids* 44, 189–197. doi: 10.1007/s00726-012-1412-x

Conflict of Interest Statement: The Associate Editor, Tunahan Cakir, declares that, despite collaborating on the Frontiers Research Topic "Endothelial cell dysfunction in pathogen-induced hemorrhagic fevers" with the author Reinhard Guthke, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 06 February 2015
doi: 10.3389/fmicb.2015.00055

Bioinformatic and mass spectrometry identification of *Anaplasma phagocytophilum* proteins translocated into host cell nuclei

Sara H. G. Sinclair[1,2,3,4], Jose C. Garcia-Garcia[2,5] and J. Stephen Dumler[1,2,3,4]*

[1] Graduate Program in Cellular and Molecular Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3] Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA
[4] Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA
[5] Procter and Gamble Co., Cincinnati, OH, USA

Edited by:
Saliha Durmus, Gebze Technical University, Turkey

Reviewed by:
Uygar Tazebay, Gebze Technical University, Turkey
Mehmet Mete Altintas, Rush University, USA

***Correspondence:**
J. Stephen Dumler, Departments of Pathology and Microbiology and Immunology, University of Maryland School of Medicine, Health Sciences Facility-1, Room 322D, 685 W. Baltimore St., Baltimore, MD 21201, USA
e-mail: sdumler@som.umaryland.edu

Obligate intracellular bacteria have an arsenal of proteins that alter host cells to establish and maintain a hospitable environment for replication. *Anaplasma phagocytophilum* secrets Ankyrin A (AnkA), via a type IV secretion system, which translocates to the nucleus of its host cell, human neutrophils. *A. phagocytophilum*-infected neutrophils have dramatically altered phenotypes in part explained by AnkA-induced transcriptional alterations. However, it is unlikely that AnkA is the sole effector to account for infection-induced transcriptional changes. We developed a simple method combining bioinformatics and iTRAQ protein profiling to identify potential bacterial-derived nuclear-translocated proteins that could impact transcriptional programming in host cells. This approach identified 50 *A. phagocytophilum* candidate genes or proteins. The encoding genes were cloned to create GFP fusion protein-expressing clones that were transfected into HEK-293T cells. We confirmed nuclear translocation of six proteins: APH_0062, RplE, Hup, APH_0382, APH_0385, and APH_0455. Of the six, APH_0455 was identified as a type IV secretion substrate and is now under investigation as a potential nucleomodulin. Additionally, application of this approach to other intracellular bacteria such as *Mycobacterium tuberculosis, Chlamydia trachomatis* and other intracellular bacteria identified multiple candidate genes to be investigated.

Keywords: ***Anaplasma phagocytophilum*, nucleomodulin, nuclear translocation, oxidative burst, iTRAQ**

INTRODUCTION

Anaplasma phagocytophilum is an obligate intracellular bacterium of human neutrophils. The neutrophil is an unlikely host as it creates an intracellular milieu that is a highly inhospitable environment for bacterial survival. Yet, *A. phagocytophilum* requires the neutrophil for propagation and survives by altering the cellular antimicrobial properties while paradoxically increasing pro-inflammatory functions (Banerjee et al., 2000; Carlyon et al., 2002; Borjesson et al., 2005; Choi et al., 2005; Carlyon and Fikrig, 2006). The fitness advantage gained with suppression of microbial killing while enhancing recruitment of new host cells for population expansion is the benefit of this paradoxical dichotomy of functional reprogramming. There is increasing evidence to suggest that the bacterium accomplishes this with coordinated reprogramming of neutrophil gene transcription by reorganizing large regions of host cell chromatin (Sinclair et al., 2014).

Importantly, *A. phagocytophilum* produces a protein, Ankyrin A (AnkA) that is exported from the bacterium and eventually localizes to the nucleus of the infected host cell (Caturegli et al., 2000; Park et al., 2004). Previously, our laboratory investigated the effect of infection on the transcriptional repression of *CYBB*, encoding $gp91^{phox}$ (Garcia-Garcia et al., 2009a,b). AnkA is capable of directly binding host cell DNA, and in the case of *CYBB*, transcription is dampened when AnkA binds to its proximal promoter (Park et al., 2004; Garcia-Garcia et al., 2009a,b). Furthermore, increased histone deacetylase (HDAC) activity enhances *A. phagocytophilum* infection in part because AnkA recruits HDAC1 to the *CYBB* promoter to close the chromatin and exclude RNA polymerase binding (Garcia-Garcia et al., 2009a; Rennoll-Bankert and Dumler, 2012). Owing to their capacity to enter the nucleus and modulate host cell transcription, microbial factors such as AnkA have been called "nucleomodulins."

It is currently unclear as to whether HDAC recruitment is the predominant mechanism by which AnkA exerts its chromatin modulating effects, whether there are other host factors (e.g., polycomb repressive or hematopoietic associated factor-1 [HAF1] complexes), or additional bacterial-derived nucleomodulins that further contribute to reprogramming. The *A. phagocytophilum* genome encodes a type 4 secretion system (T4SS) that allows the bacteria to translocate effector proteins into the host cytosol (Dunning Hotopp et al., 2006; Lin et al., 2007; Rikihisa et al., 2010). AnkA was the first T4SS substrate identified among Rickettsiales, and it plays a critical and potentially dominant role

in the course of establishing and sustaining neutrophil infection (IJdo et al., 2007; Garcia-Garcia et al., 2009a,b; Al-Khedery et al., 2012; Rennoll-Bankert and Dumler, 2012). In contrast to other gram-negative T4SSs, the *vir* genes encoding the secretion system of the *Rickettsiales* family are organized differently in that they are clustered in three different genomic locations (Ohashi et al., 2002; Rikihisa et al., 2010). Between individual *A. phagocytophilum* strains, variations of the T4SS appear to contribute to host specificity and strain virulence (Al-Khedery et al., 2012).

We hypothesize that *A. phagocytophilum* expresses additional nuclear effector proteins secreted by its type IV secretion system (T4SS) and that these also play a role in pathogenicity. It is likely that some will be nucleomodulins which could contribute to transcriptional reprogramming of infected neutrophils. Pilot studies using bioinformatics tools, and iTRAQ protein profiling among infected and uninfected cells were used to identify candidate proteins that potentially localize to the host cell nucleus. The profiling identified 50 *A. phagocytophilum* proteins, one of which was AnkA, and at least 7 of these were predicted to enter the nucleus based on the presence of both a nuclear localization sequence and a bacterial secretion signal sequence. Ultimately, 3 of the 7 proteins identified in the bioinformatic screen and 3 of 37 identified by iTRAQ profiling of nuclei from infected cells translocated into HEK-293T human embryonic kidney and PLB-985 granulocytic cell nuclei.

METHODS

IN SILICO PREDICTION OF *A. PHAGOCYTOPHILUM* PROTEINS TARGETED TO THE HOST CELL NUCLEUS

Our initial focus was on proteins involved in regulation of host gene expression. Since these events occur chiefly in the nucleus, we developed an unbiased computational approach to identify potential nucleomodulins encoded in intracellular bacterial genomes based on their likelihood for translocation into the host cell nucleus and applied this to the *A. phagocytophilum* HZ strain genome (Supplemental Figure 1). Annotated protein tables for bacteria, focusing on the *A. phagocytophilum* HZ strain genome, were obtained from the National Center for Biotechnology Information (NCBI) database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The *A. phagocytophilum* protein table was used as the database for eukaryotic subcellular localization search algorithms. Although we used a database with 1264 annotated *A. phagocytophilum* proteins, including hypothetical proteins, multiple programs were implemented to obtain high prediction accuracy and processing capacity. Since we needed only to predict nuclear proteins, localization coverage was not taken into account. Since hybrid methods are preferable when little is known about the protein of interest (Donnes and Hoglund, 2004) we used ProtComp Version 6 (Softberry, Inc.), a computational algorithm for the identification of sub-cellular localization of eukaryotic proteins.

We next applied PSORTb v.2.0 to exclude potential membrane proteins that are unlikely to be secreted into the host cell (Gardy et al., 2005). Finally, we used computational algorithms to predict the presence of eukaryotic nuclear localization signals (NLS). NLSs often possess sequences with a high basic amino-acid content (Hicks and Raikhel, 1995) and are generally classified into three categories: classical or monopartite (NLSm), bipartite (NLSb), and a type of N-terminal signal found in yeast protein, Mat alpha2, a poorly studied signal that is not incorporated in most NLS prediction algorithms. To screen broadly for potential NLSs, we selected MultiLoc (Hoglund et al., 2006). MultiLoc also identifies matches in NLSdb, a database of experimentally known NLSs (Nair et al., 2003) and is also useful to predict NLSm and NLSb in addition to the NLSdb attribute, since NLSdb recognizes only 43% of the nuclear proteins. MultiLoc calculates a probability estimate for each subcellular location and the protein is assigned to the compartment with the highest score. The MultiLoc output was recorded and used to calculate a Nuclear Score that better reflects the purpose of the search:

$$\text{Nuclear Score} = \text{MultiLoc Nuc} + \text{NLSdb} + (0.5 \times \text{NLSm} + 0.5 \times \text{NLSb})$$

where: (i) $0 \leq$ MultiLoc Nuc <1 is the probability estimate of the protein being nuclear as calculated by MultiLoc; (ii) NLSdb is 1 if the protein contains a known NLS, 0 if not; (iii) NLSm is 1 if the protein contains a predicted NLSm, 0 if not; and (iv) NLSb is 1 if the protein contains a predicted NLSb, 0 if not. Weighting was applied since the presence of a predicted NLS suggests, but is not conclusive; therefore the NLSm or NLSb prediction contributes only half to the final nuclear score. The addition of the continuous MultiLoc Nuc score provides a better ranking of the proteins, given that the other indicators contribute discretely to the Nuclear Score. However, no proteins without a known or predicted NLS will produce a Nuclear Score >1 since MultiLoc Nuc <1.

iTRAQ FOR IDENTIFICATION OF POTENTIAL NUCLEAR-TRANSLOCATED PROTEIN PROFILING

A. phagocytophilum-infected and uninfected HL-60 cells, a promyelocytic cell line commonly used for *A. phagocytophilum* propagation as previously described (Goodman et al., 1996; Park et al., 2004), were fractionated to obtain nuclei and nuclear proteins. iTRAQ (isobaric tag for relative and absolute quantitation protein profiling technology [Applied Biosystems]), a mass spectrometric technique where 2 protein expression profiles are compared, was used to identify candidate bacterial proteins present in the nucleus of infected cells. One hundred μg in replicate samples from nuclear fractions of infected and uninfected HL-60 cells were acetone-precipitated and checked for protein integrity and sample quality. The samples were reduced and cysteines blocked following the iTRAQ kit protocol (Applied Biosystems). Samples were digested with trypsin overnight at 37°C and then labeled with iTRAQ tags in replicates, pooled and fractionated using a strong cation exchange (SCX) column on an Ultimate HPLC system (LC Packings). Approximately 20 fractions were collected and analyzed on Qstar Pulsar™ (Applied Biosystems-MDS Sciex) interfaced with an Agilent 1100 HPLC system. Peptides were separated on a reverse-phase column, and MS/MS analysis was performed. The MS/MS spectral data were extracted and searched against Uniprot-sprot database (entries for *Homo sapiens* and *A. phagocytophilum*) using ProteinPilot™ software (Applied Biosystems). For each protein, two types of

scores were reported: unused ProtScore and total ProtScore. The total ProtScore is a measurement of all the peptide evidence for a protein and is analogous to protein scores reported by other protein identification software programs. However, the unused ProtScore is a measurement of all the peptide evidence for a protein that is not better explained by a higher ranking protein and was the method of choice. The protein confidence threshold cut-off for this study was set at an unused score of 2.0 with at least one peptide with 99% confidence. A ratio of infected to uninfected (Aph:HL-60) score was used to identify *A. phagocytophilum* proteins in nuclear lysates. To do this, we averaged the ratios of uninfected HL-60 nuclear lysate replicates (isobaric isotope labels 115:114) and ratios of nuclear lysate replicates from *A. phagocytophilum*-infected HL-60 cells (116:114 and 117:114) to create the composite Aph:HL-60 mean ratio. Proteins identified with mean ratios (infected/uninfected) >1.2 were selected for further study.

GFP-FUSION PROTEIN PLASMID CLONES AND TRANSFECTIONS

Forty one GFP C-terminal fusion proteins were prepared using pMAXFP-Green-C (Lonza, cat# AMA-VDF1011) and the Infusion HD Liquid cloning kit (Clontech). Briefly, target genes were amplified using PlatinumTaq (Life Technologies) and PCR purified using Qiagen PCR purification kits (Qiagen). Primers were designed using Clontech's Online Infusion tools, Primer Design. Amplicons were created to be fused with the pMAXFP-Green-C vector after digestion with *Xho*I. Primers were approximately 40–45 bp in length and had the sequence GAAGAAAGATCTCGAGCT added to the 5′ end of the forward gene-specific primer (20–25 bp), and GAAGCTTGAGCTCGAGT added 5′ to the reverse primer (Supplemental Table 1). The Infusion-HD kit instructions were followed as per the manufacturer's recommendations. Clones were transformed into *E. coli* JM109 (Promega) and after antibiotic selection, were sequenced to ensure they were in the correct orientation and in frame. HEK-293T cells were transfected with GFP-fusion vectors using Lipofectamine 2000 (Life Technologies) and PLB-985 cells were transfected with the Amaxa Nucleofector shuttle and the SF kit reagents (Lonza) as per manufacturer's recommendations. PLB-985 cells are human myelomonoblast leukemia cells that easily differentiate into neutrophil-like cells and are readily transfected as opposed to HL-60 cells, a common host cell model for *A. phagocytophilum* infection (Pedruzzi et al., 2002; Ellison et al., 2012; Rennoll-Bankert et al., 2014). Cells were stained with DAPI 24 h later and imaged by fluorescence microscopy, gathering both superimposed green fluorescent protein and DAPI images.

DETERMINATION OF T4SS SUBSTRATES

A. phagocytophilum proteins that localized to the nucleus of HEK-293T and PLB-985 cells were tested for their ability to be secreted by the T4SS Dot/Icm system of *Coxiella burnetii* RSA439 avirulent phase II nine-mile strain using adenylate cyclase translocation assays (Larson et al., 2013). Fusion proteins were created by cloning full-length coding regions or C-terminal 100 aa truncations to the *Bordetella pertussis* adenylate cyclase gene (*cyaA*). To achieve this, *C. burnetii* was transiently propagated in ACCM-2 axenic culture medium and transformed with the constructs (performed at the NIAID Rocky Mountain Laboratories [Hamilton, MT] by Paul Beare, Ph.D. and Charles L. Larson). Axenic *C. burnetii* was transformed by electroporation and cultured in ACCM-2 medium for 24 h followed by chloramphenicol selection (Beare et al., 2009; Voth et al., 2011). The ability of the constructs to be secreted by the Dot/Icm system was determined by measuring changes in intracellular cAMP levels. CyaA fusion proteins that contain a T4SS signal are capable of being secreted and mediate a measurable increase in cAMP. *C. burnetii* transformants containing the *cyaA* constructs were used to infect THP-1 cells (a human myelomonocytic cell line) at an MOI of 100:1, and included both *A. phagocytophilum* AnkA (APH_0740) and *Coxiella* vacuolar protein A (CvpA), both known T4SS substrates as positive controls. After 3 days, the cells were harvested, lysed and examined for cAMP production by enzyme immunoassay. Results were expressed as fold change in intracellular cAMP concentration compared to empty vector control (CyaA only) that lacked a T4SS signal; values >2 were considered positive for type 4 secretion; values between 1 and 2 were considered marginal.

ASSAY FOR DETECTION OF REACTIVE OXYGEN SPECIES

Superoxide production was detected as described previously (Rennoll-Bankert et al., 2014). Briefly, HL-60 cells were incubated with 0.25 mM 2′,7′-dichlorofluoresein diacetate (DCFH-DA) in PBS for 30 min at room temperature. 10^5 cells were stimulated in triplicate with 1 μg/mL phorbol 12-myristate 12-acetate (PMA) and fluorescence was measured every 2 min. The relative fluorescence units at 180 min were averaged and compared to unstimulated controls using a two-sided Student's *t*-test, α 0.05.

RESULTS

IN SILICO PREDICTION OF *A. PHAGOCYTOPHILUM* PROTEINS TARGETED TO THE HOST CELL NUCLEUS

Of 1264 proteins and hypothetical proteins examined by the bioinformatics algorithm, 123 were identified by ProtComp as nuclear-localized; 3 of these were classified in PSORTb as potentially nuclear membrane-associated; after analysis of NLSdb and screening for NLSm and NLSb, 7 candidate proteins had a total Nuclear score >1 (**Table 1**). One candidate with a high ProtComp score for nuclear localization but that lacked a predicted NLS (APH_0805) was selected as a control. The known nuclear-translocated AnkA was not identified in this screen.

iTRAQ IDENTIFICATION OF *A. PHAGOCYTOPHILUM* NUCLEAR-TRANSLOCATED PROTEINS

We detected 43 *A. phagocytophilum* proteins with an Aph:HL-60 ratio >1.2 in the nucleus of infected cells (**Table 2**), including the top hit, AnkA that is established to translocate into the nucleus. This approach allowed the identification of *A. phagocytophilum* proteins most likely to have been translocated into the nucleus and provided a more complete list of candidates to investigate out of the 1264 *A. phagocytophilum* ORFs available for study. Of these 43 candidates, only AnkA was excluded from subsequent cloning and expression for *in vitro* nuclear localization studies.

IN VITRO NUCLEAR LOCALIZATION

As an inclusive screen, and because contamination of nuclear preparations could not be entirely excluded in iTRAQ studies, nuclear localization of proteins identified by bioinformatic

Table 1 | Bioinformatic prediction of *A. phagocytophilum* nuclear-translocated proteins, by likelihood based on Final Score rank.

Locus name	Acc. No.	Protein	Gene	Ranking	ProtComp	MultiLoc	NLS	NLS1	NLS2	Final score
APH_0820[a]	YP_505397.1	Hypothetical protein		10	2.1	0.94	1	1	0	2.44
APH_0847	YP_505424.1	Hypothetical protein		22	2.1	0.97	0	1	1	1.97
APH_0382	YP_504988.1	HGE-14 protein		56	1.7	0.97	0	1	0	1.47
APH_0385	YP_504990.1	HGE-14 protein		75	2.1	0.94	0	1	0	1.44
APH_0455	YP_505057.1	HGE-14 protein		76	2.2	0.94	0	1	0	1.44
APH_0485	YP_505084.1	Hypothetical protein		77	2.2	0.94	0	1	0	1.44
APH_0576	YP_505167.1	RNA polymerase sigma factor RpoD	*rpoD*	114	2.1	0.89	0	1	0	1.39
APH_0805[b]	YP_505382.1	Hypothetical protein		1891	2.1	0.96	0	0	0	0.96

[a] *Not cloned.*

[b] *Selected as negative control.*

methods or by iTRAQ mass spectrometry were confirmed by cloning the corresponding genes into a mammalian expression vector for expression as GFP fusion proteins. APH_0805 that was predicted to have nuclear localization yet lacked a predicted NLS and had a below-threshold Nuclear score was used as a non-translocating control. HEK-293T cells and PLB-985, a promyelocytic cell line, were transfected and examined for nuclear localization of the GFP-fusion proteins with Hoescht 33342 nuclear counterstaining. Six of the 42 proteins tested (36 from iTRAQ profiling, 7 from the bioinformatic screen), translocated to the nucleus: APH_0062 (hypothetical protein), RplE (50S ribosomal protein L5 [APH_0292]), Hup (DNA-binding protein HU [APH_0783]), and APH_0455, APH_0382, and APH_0385 (all HGE-14) (**Figure 1** and Supplemental Figure 2). APH_0278 (*tuf-1*; elongation factor Tu) was not cloned, but instead the identical APH_1032 (*tuf-2*; elongation factor Tu) was used but did not enter the nucleus. Nine proteins were either unable to be cloned or cloning was not attempted, including: APH_0160 (putative thymidylate synthase, flavin-dependent, truncation, partial); APH_0196 (nitrogen assimilation regulatory protein); APH_0289 (ribosomal protein S17 [*rpsQ*]); APH_0820 (hypothetical protein); APH_0906 (hypothetical protein); APH_1023 (DNA-directed RNA polymerase, beta subunit [*rpoC*]); APH_1024 (DNA-directed RNA polymerase, beta subunit [*rpoB*]); APH_1034 (ribosomal protein S7 [*rpsG*]) and APH_1333 (transcription elongation factor GreA).

DETERMINATION OF TYPE 4 SECRETION SUBSTRATES

Proteins identified to localize to the nucleus were further investigated to determine if they could be secreted by the T4SS of *Coxiella burnetii*, which is similar to that of *A. phagocytophilum*. T4SS substrate status was determined by the ability of the CyaA-fusion to exit *C. burnetii* and produce a measurable increase in cAMP concentrations with infection of THP-1 cells. Of the 6 genes tested only APH_0455 was identified to be a type 4 secretion substrate (**Figure 2**).

DETERMINATION OF OXIDATIVE BURST AFTER TRANSFECTION AND NUCLEAR TRANSLOCATION

GFP-fusion constructs were transfected into HL-60 cells to determine their ability to alter the oxidative burst response. Unfortunately, the methods (electroporation, lipofectamine, viral transduction) used to transfect the HL-60 cells (differentiated or undifferentiated), abrogated oxidative burst as compared with non-transfected cells. Thus, we compared results to PMA-stimulated oxidative burst in HL-60 cells transfected with the empty GFP plasmid as control. When RFU values of each unstimulated transfected control cell culture were compared to PMA-stimulated, significant oxidative burst, as seen with the GFP plasmid control, was observed only with Hup and APH_0382 (**Figure 3A**). When normalized to GFP plasmid transfection alone, APH_0062, RplE, APH_0455, Hup, and APH_0385 significantly repressed respiratory burst (**Figure 3B**). However, responses varied in intensity over several repeated experiments, likely in part due to the variable transfection efficiency obtained with HL-60 cells. These data suggest that one or more of these effectors could contribute to dampened production of reactive oxygen species.

DISCUSSION

While considerable focus has been placed on AnkA as the primary nucleomodulin of *A. phagocytophilum*, it does not seem plausible that a single protein can account for the widespread transcriptional and phenotypic changes induced with infection. Using current bioinformatics tools and mass spectrometry, a number of other proteins encoded in the *A. phagocytophilum* genome were identified that could potentially localize to the host cell nucleus. To validate the candidate genes, GFP-fusion proteins were created and screened for nuclear localization within HEK-293T cells. This approach narrowed the list of target genes for further investigation to six.

No candidate proteins were identified in both the bioinformatic screen and in the iTRAQ mass spectrometry analysis. If one assumes that the mass spectrometry data is accurate, the bioinformatic approach was ineffective at identifying features to predict nuclear localization for 3 of the six proteins shown capable of entering the nucleus; as a result, APH_0062 (cytoplasmic), *hup*, and *rplE* (both mitochondrial) were excluded from the bioinformatic identification because they were not assigned a nuclear localization. In contrast, no bioinformatic-predicted candidate appeared in the iTRAQ mass spectrometry analyses, suggesting limitations in sensitivity and/or contamination of nuclear preparations by non-nuclear localized proteins. Thus, the combination of both approaches increased the ability to identify

Table 2 | *Anaplasma phagocytophilum* proteins identified in the nuclear lysates of infected HL-60 cells by iTRAQ with ratios compared with uninfected cells of >1.2 and ranked by Unused ProtScore to identify high likelihood candidates for nuclear translocation.

Locus name	Accession	Protein	Gene	Ratios of labeled peptides[a]			Mean HL-60	Mean Aph	Ratio Aph:HL-60	Unused ProtScore
				115:114	116:114	117:114				
APH_0740	gi|88607707	Ankyrin A	*ankA*	1.06	1.42	1.43	1.03	1.43	1.38	40.10
APH_1023	gi|88607105	DNA-directed RNA polymerase subunit beta; RNAP subunit beta	*rpoC*	1.04	1.50	1.39	1.02	1.45	1.42	32.10
APH_0240	gi|88606723	60 kDa chaperonin GroEL	*groEL*	1.00	1.88	1.93	1.00	1.90	1.90	30.60
APH_1024[2]	gi|88606872	DNA-directed RNA polymerase subunit beta; RNAP subunit beta	*rpoB*	1.02	1.33	1.27	1.01	1.30	1.28	23.40
APH_0906	gi|88606911	Hypothetical protein APH_0906		1.05	1.25	1.22	1.02	1.24	1.21	20.70
APH_0278[2]	gi|88607578	Translation elongation factor Tu; EF-Tu	*tuf1*	1.18	1.93	1.83	1.09	1.88	1.73	20.20
APH_1099	gi|88607685	DNA-binding response regulator CtrA	*ctrA*	1.05	2.65	2.71	1.03	2.68	2.62	19.90
APH_0303	gi|88606699	DNA-directed RNA polymerase subunit alpha; RNAP subunit alpha	*rpoA*	1.16	1.91	1.84	1.08	1.88	1.74	15.90
APH_0784	gi|88606926	DNA-binding protein HU	*hup*	1.00	2.38	2.03	1.00	2.21	2.21	15.40
APH_0968	gi|88606840	ATP-dependent protease La	*lon*	1.01	1.62	1.44	1.01	1.53	1.52	15.10
APH_1100	gi|88606714	DNA-binding protein		0.99	2.82	2.44	0.99	2.63	2.64	13.80
APH_0469	gi|88607025	Putative malonyl-CoA decarboxylase		1.04	1.27	1.28	1.02	1.27	1.24	12.60
APH_0445	gi|88607683	Transcription elongation factor NusA	*nusA*	1.00	1.53	1.50	1.00	1.52	1.52	12.20
APH_0339	gi|88607311	Putative thermostable metallocarboxypeptidase		1.11	1.59	1.58	1.05	1.58	1.50	9.60
APH_1239	gi|88607921	P44–15b outer membrane protein; major surface protein-2C	*p44–15b*	1.05	3.60	3.63	1.03	3.62	3.53	9.10
APH_0062	gi|88606901	Hypothetical protein APH_0062		1.06	1.91	1.77	1.03	1.84	1.79	8.70
APH_1097	gi|88607712	DNA polymerase III, beta subunit	*dnaN*	1.14	1.33	1.30	1.07	1.32	1.23	6.60
APH_0135	gi|88606701	Cold shock protein, CSD family		0.97	1.79	1.78	0.99	1.79	1.81	6.30
APH_0397	gi|88606909	30S ribosomal protein S2	*rpsB*	1.07	1.53	1.45	1.04	1.49	1.44	6.20
APH_1263	gi|88607227	Translation initiation factor IF-3	*infC*	0.94	2.12	1.97	0.97	2.05	2.11	5.00
APH_0398	gi|88607503	Elongation factor Ts; EF-Ts	*tsf*	1.03	1.24	1.28	1.01	1.26	1.24	4.40
APH_1151	gi|88607101	Hypothetical protein APH_1151		1.20	2.16	2.11	1.10	2.13	1.94	4.10
APH_0288	gi|88607038	50S ribosomal protein L29	*rpmC*	1.03	1.77	1.66	1.01	1.72	1.69	4.10
APH_1029	gi|88607731	Transcription termination/antitermination factor NusG	*nusG*	0.94	1.65	1.72	0.97	1.69	1.74	4.00
APH_1027	gi|88607420	50S ribosomal protein L1	*rplA*	1.15	1.37	1.26	1.08	1.31	1.22	3.30
APH_0515	gi|88606905	Expression regulator ApxR	*apxR*	0.99	2.02	1.94	0.99	1.98	2.00	3.20
APH_0097	gi|88606982	Protein-export protein SecB	*secB*	1.11	1.81	1.67	1.06	1.74	1.64	3.00
APH_0292	gi|88606711	50S ribosomal protein L5	*rplE*	1.26	1.52	1.57	1.13	1.54	1.36	2.40
APH_0106	gi|88607568	Riboflavin synthase, alpha subunit	*ribE*	0.98	2.04	2.00	0.99	2.02	2.04	2.30
APH_0280	gi|88607449	50S ribosomal protein L3	*rplC*	1.07	1.60	1.64	1.04	1.62	1.56	2.30
APH_0629	gi|88607793	Malate dehydrogenase	*mdh*	0.98	1.25	1.21	0.99	1.23	1.24	2.30
APH_0160[2]	gi|88606875	Putative thymidylate synthase, flavin-dependent, truncation		1.14	1.50	1.37	1.07	1.44	1.34	2.20
APH_0154	gi|88607134	Serine hydroxymethyltransferase SHMT	*glyA*	1.06	1.29	1.26	1.03	1.27	1.24	2.10
APH_0971	gi|88607838	Trigger factor; TF	*tig*	1.04	1.37	1.28	1.02	1.32	1.30	2.00
APH_0659	gi|88607183	Antioxidant, AhpC/Tsa family		0.99	1.26	1.23	1.00	1.24	1.25	2.00
APH_1349	gi|88606948	Glyceraldehyde-3-phosphate dehydrogenase, type I	*gap*	0.86	1.13	1.19	0.93	1.16	1.24	2.00
APH_1198	gi|88606994	2-oxoglutarate dehydrogenase, E2 component, dihydrolipoamide succinyltransferase	*sucB*	0.98	1.24	1.16	0.99	1.20	1.21	2.00
APH_1025	gi|88607605	50S ribosomal protein L7/L12	*rplL*	1.01	1.85	1.79	1.01	1.82	1.81	1.90

(Continued)

Table 2 | Continued

Locus name	Accession	Protein	Gene	Ratios of labeled peptides[a]			Mean HL-60	Mean Aph	Ratio Aph:HL-60	Unused ProtScore
				115:114	116:114	117:114				
APH_0289[2]	gi|88607574	30S ribosomal protein S17	*rpsQ*	0.90	1.57	1.46	0.95	1.52	1.60	1.70
APH_1034[2]	gi|88607212	30S ribosomal protein S7	*rpsG*	1.16	1.42	1.41	1.08	1.41	1.31	1.50
APH_0196[2]	gi|88607673	Response Regulator NtrX, putative nitrogen assimilation regulatory protein	*ntrx*	0.93	1.70	1.50	0.97	1.60	1.66	1.40
APH_1333[2]	gi|88607617	Transcription elongation factor GreA	*greA*	0.95	1.29	1.27	0.98	1.28	1.31	1.30
APH_1098	gi|88607131	3′–5′ exonuclease family protein		1.12	1.30	1.29	1.06	1.30	1.22	1.30

[a] *Isobaric ion labels of nuclear lysates from: 114 and 115, uninfected HL-60 cells; 116 and 117, A. phagocytophilum-infected HL-60 cells.*

[b] *Not cloned.*

and exclude candidates for further analysis. It is important to note that the screen will only identify those genes capable of entering the nucleus on their own accord via an identified or unidentified nuclear localization signal. Some proteins identified as present in the nucleus in the iTRAQ screen could indeed localize to the nucleus but might not be confirmed by transfection screens. A bacterial-derived protein shuttled into the nucleus as a component of a protein complex, or one that possesses an uncharacterized NLS, as is the case with AnkA, would not be identified. Furthermore, HEK-293T cells are not a model cell line for *A. phagocytophilum* infection and transfection of these proteins does not mimic infection, a much more complex process; therefore, confirmation of nuclear translocation in PLB-985 was performed.

Additionally, *A. phagocytophilum* is largely refractory to gene delivery by genetic transformation. Previous reports demonstrate *A. phagocytophilum* transformation using the Himar1 transposase system that introduces small GFP proteins or disrupts bacterial genes and consequently protein expression (Felsheim et al., 2006; Chen et al., 2012). This process does not result in gene entry, but results in a library of mutant bacteria that can facilitate complex functional studies and insight into the importance of mutated genes for establishing or maintaining infection. However, directed mutation by homologous recombination has not yet been described for *A. phagocytophilum*. None-the-less, this relatively simple experiment yielded multiple candidate genes of interest for further investigation.

After narrowing the initial bioinformatic and iTRAQ list of candidate genes to six, we investigated the ability of these proteins to be secreted by the bacterium. For *A. phagocytophilum*, the most well characterized secretion mechanism is that of the T4SS. Because of this, we focused on whether or not these proteins could be secreted by a T4SS. As an obligate intracellular bacterium that resides solely in membrane-bound vacuoles of its host cells, *A. phagocytophilum*-secreted proteins very likely first enter the cytosol before translocation to the nucleus, but are unlikely to be detected outside of the host cell owing to the intracellular vacuolar membranes accessible to the bacterium. Thus, the *C. burnetii* Dot/Icm T4SS was used as a surrogate delivery system because *C. burnetii* is capable of being transfected easily when cultivated in axenic medium but resides within host cell vacuoles when cultivated in mammalian cells. The *Dot/Icm* secretion system is compatible with that of *A. phagocytophilum* and, unless cultivated in specific axenic medium, *C. burnetii* is also an obligate intracellular bacterium residing within membrane-bound vacuoles. Using fusions with *B. pertussis* CyaA, one of six *A. phagocytophilum* candidate nuclear-localizing proteins was identified as a T4SS substrate. The remaining 5 did not appear to be secreted by the *Dot/Icm* system. Using SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/), we determined the presence of putative Sec1 secretion signals in the genes encoding APH_0382, APH_0385, and APH_0455 (all HGE-14-like); experimental confirmation of this secretion mechanism was not further attempted.

Interestingly, APH_0382, APH_0385, and APH_0455 were shown to be differentially expressed between mammalian and tick cells. The transcription of each of these proteins was approximately 2.9–3.3-fold greater in HL-60 cells than ISE6 (tick) cells (Nelson et al., 2008). This suggests that these HGE-14-like proteins likely play a role in establishing or maintaining infection in mammalian cells. In fact, differential transcription of *A. phagocytophilum* genes plays a role in the life cycle of the bacterium in mammalian and tick cells (Wang et al., 2007; Nelson et al., 2008; Troese et al., 2011; Mastronunzio et al., 2012). APH_0784 (DNA binding protein HU), and APH_0292 (50S ribosomal protein L5) are among the 20 most abundant proteins expressed in infected *I. scapularis* salivary glands (Mastronunzio et al., 2012), and both were found in nuclear lysates of infected HL-60 cells, yet predicted to localize to the mitochondrion and cytosol, respectively. We also identified the transcriptional regulator of p44/*msp2* genes, ApxR (APH_0515; Wang et al., 2007) in nuclear lysates from *A. phagocytophilum* infected HL-60 cells, but at a low unused ProtScore. As ApxR was unable to translocate to the nuclei of HEK293 cells, its presence indicates the potential for low level cytoplasmic contamination in the nuclear preparations. However, the overall level of cytoplasmic contamination is likely to be low since the most abundant *A. phagocytophilum* proteins in the P44/Msp2 family (Wang et al., 2007; Nelson et al., 2008; Mastronunzio et al., 2012) were not abundant in nuclear lysates. Finally, APH_1235 is characterized as a specific marker of dense core infectious *A. phagocytophilum* (Troese et al., 2011). It is among the 20 most abundantly-expressed proteins in tick salivary glands (Mastronunzio et al., 2012), is significantly upregulated in

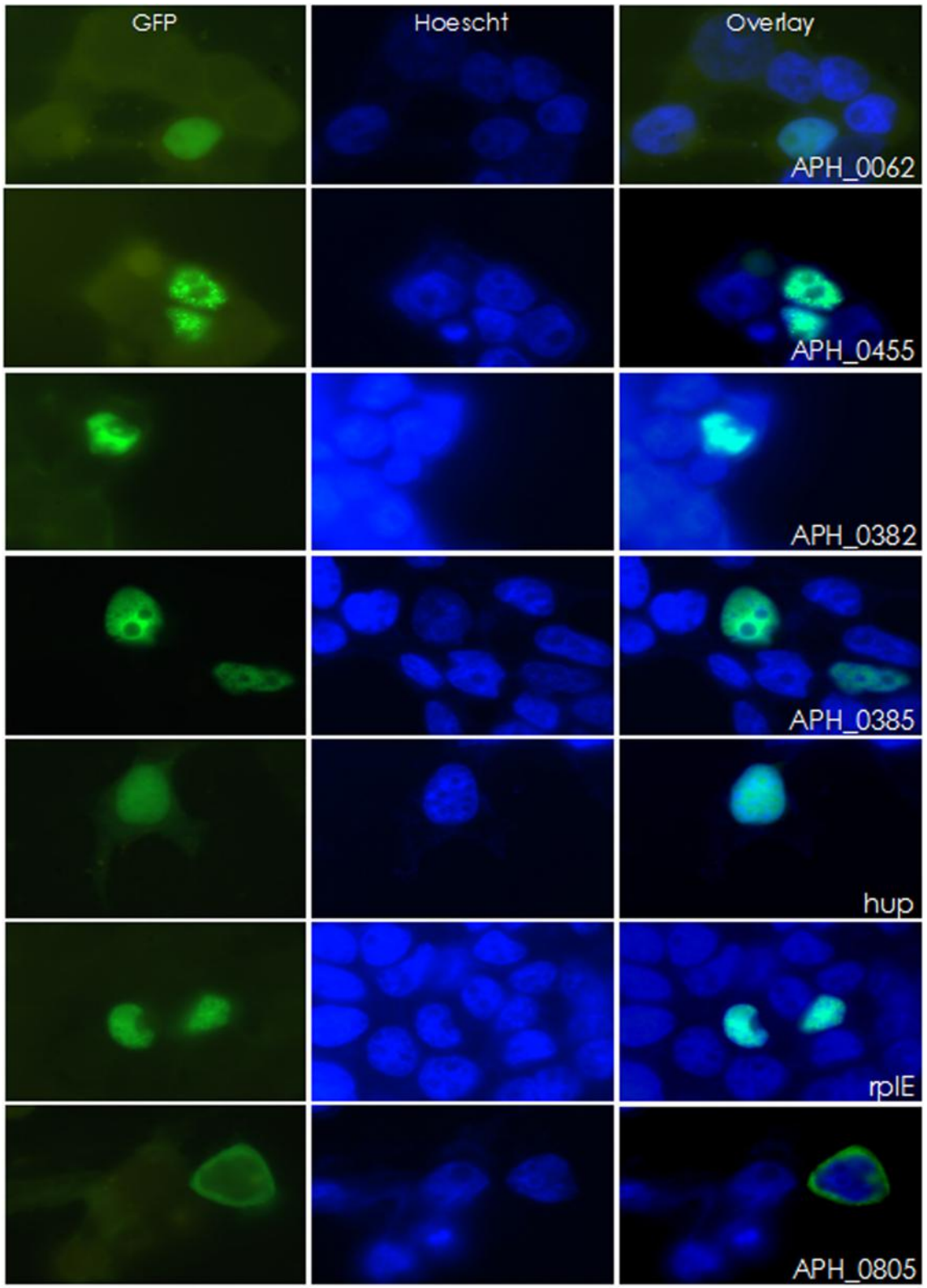


FIGURE 1 | Six *A. phagocytophilum* candidate genes were found to localize to the nucleus of HEK-293T cells. Candidate genes were fused to GFP and transfected into HEK-293T cells. Twenty four hours post-transfection, cells were stained with DAPI and imaged. Of the 42 GFP-fusion proteins created, six localized to the nucleus. APH_0805 is shown here as an example of a protein that did not localize to the nucleus.

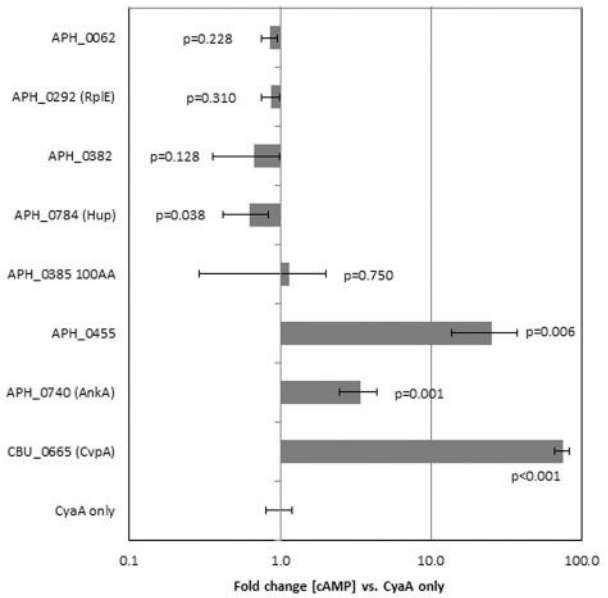


FIGURE 2 | APH_0455 is secreted by the *Dot/Icm* T4SS of *C. burnetii*. Candidate genes were fused to *B. pertussis* adenylate cyclase (*cyaA*), transfected into *C. burnetii* and selected by chloramphenicol resistance. Transformed *C. burnetii* clones were then used to infect THP-1 cells. Three days post-transfection, THP-1 cells were assayed for cAMP production. Only those constructs that contain a T4SS signal sequence have measurable changes in cAMP production. The results represent the average of two separate experiments each with replicate tests. The *p*-values were calculated based on comparisons with fold change of *C. burnetii* transformed by empty plasmid (CyaA only) using two-sided Student's *t*-tests, $\alpha = 0.05$. CBU_0655 is CvpA, a known T4SS substrate of *C. burnetii*.

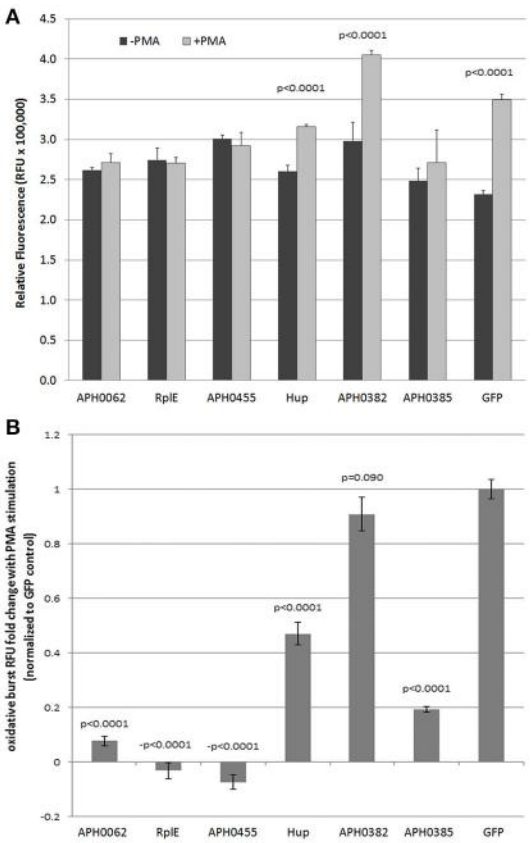


FIGURE 3 | Expression of putative nuclear effectors APH_0062, RplE and APH_0455 dampen PMA-stimulated reactive oxygen species production by HL-60 cells. HL-60 cells were transfected with 2 μg plasmid and assayed for respiratory burst 48 h later. **(A)** The average of three replicates is displayed ±SEM at 180 min. **(B)** The fold change was calculated by dividing the ratio of PMA stimulation of each transfectant to the GFP control. *P*-values were calculated using Students *t*-tests.

dense core cells with HL-60 cell infection (Troese et al., 2011), and is believed to facilitate tick to mammal transmission. While predicted to localize to the nucleus by ProtComp v.6 and identified in infected HL-60 cell nuclear lysates, published data demonstrate the lack of nuclear localization (Troese et al., 2011). Moreover, it lacked a recognized NLS and the iTRAQ unused score was low, suggesting low-level contamination from the host cytosol.

APH_0455, a HGE-14 protein, is of particular interest owing to its utilization of the T4SS to enter the cell and its translocation into the nucleus where it forms small aggregates and clusters dispersed unevenly throughout the nucleoplasm. APH-0455 is one of several HGE-14 proteins predicted to enter the nucleus, and APH-0455 has been described to have transmembrane domains that would predict it to be a type II membrane protein, and possesses 4 conserved 41 amino acid repeats followed by 2 similar truncated repeats (Lodes et al., 2001). This repeat region overlaps a region with a conserved Med15/ARC15 (pfam09606) domain. Med15/ARC105 domains are found as part of a family of sterol regulatory element binding proteins (SREBPs), transcription activators that regulate genes involved in cholesterol and fatty acid homeostasis. In humans, SREBPs bind CREB-binding protein (CBP)/p300 acetyltransferase that in turn affect chromatin structure and gene transcription (Yang et al., 2006). Whether APH_0455 plays a role in these critical pathways for *A. phagocytophilum* survival needs to be determined (Lin and Rikihisa, 2003).

Because of the candidate proteins' abilities to act as T4SS or Sec1 substrates and to localize to the nucleus, we sought to determine if they played a role in altering the phenotype of HL-60 cells, a commonly used cell model for *A. phagocytophilum* infection. Unfortunately, transfection of HL-60 cells with a variety of methods inconsistently altered oxidative burst capacity, and often the vehicle controls and transfection reagents were enough to abrogate responses. Despite the variable responses, we observed trends toward reduction of oxidative burst (**Figure 3**). Despite these trends, we cannot currently conclude with certainty that these effectors play a role in limiting oxidative burst as shown for AnkA (Banerjee et al., 2000).

For each of the *A. phagocytophilum* proteins that localized to the nucleus of HEK-293T and PLB-985 cells, it would be important confirm their presence in the nuclei of *A. phagocytophilum*-infected cells visually or biochemically, and to potentially assess the effects of their absence in *A. phagocytophilum* among Himar1 transposase libraries (Nelson et al., 2008; Troese et al.,

2011). Additionally, future studies will examine their role in transcriptional and functional changes in differentiated HL-60 cells, the preferred model for *A. phagocytophilum*-directed neutrophil reprogramming. Such studies will focus on transcriptional responses, functional assays and, given the role that AnkA plays during the course of infection, studies of nuclear protein-protein, DNA-protein, and RNA-protein interactions. The screening techniques modeled here using *A. phagocytophilum* will allow for a more focused approach to identify potential nucleomodulins and could facilitate studies of microbial nucleomodulin manipulation of host cell transcriptional programs.

These techniques are not limited to the *A. phagocytophilum* genome but can also be applied to other intracellular bacteria. Using the same bioinformatics approaches (Supplemental Methods, Supplemental Figure 1, Supplemental Tables 1, 2), candidate genes were identified for other pathogens including, but not limited to: *Chlamydia trachomatis, Coxiella burnetii, Ehrlichia chaffeensis, Mycobacterium tuberculosis, Yersinia pestis, Legionella pneumophila, Francisella tularensis,* and *Listeria monocytogenes*. Identification of new nucleomodulins in any one of these pathogens could add further insight as to how bacteria modulate their host cells and cause aberrant transcriptional reprogramming.

CONCLUSION

We used a combination of bioinformatic screens and iTRAQ *in vitro* identification of potential nuclear-translocated proteins to stratify and rapidly identify candidate nucleomodulins in *A. phagocytophilum*, an approach easily applied to other intracellular pathogens. By combining data gathered from bioinformatics prediction tools and iTRAQ, 50 *A. phagocytophilum* proteins were identified as potential nucleomodulins. Of the 50, we confirmed that six proteins were capable of localizing to the nucleus on their own, including APH_0455 that is also a T4SS substrate. The identification of novel nuclear translocated proteins provides additional support for the concept of nucleomodulin-mediated reprogramming of cellular functions that improve microbial fitness by promoting extended intracellular survival and more opportunities for transmission.

ACKNOWLEDGMENTS

Supported by grant R01AI044102 (J. Stephen Dumler) from the US National Institutes of Allergy and Infectious Diseases, National Institutes of Health, and by subaward PO #SR00000759 (to Jose C. Garcia-Garcia) from Region III Mid-Atlantic Regional Center for Excellence grant U54 AI057168 from the NIAID (M. M. Levine, PI). The authors would like thank Paul Beare, Ph.D. and Charles Larson (Rocky Mountain Laboratories, NIAID, Hamilton, MT USA) for advice and performance of the T4SS assays; Kristen Rennoll-Bankert, Ph.D. (University of Maryland School of Medicine, Baltimore, MD USA) for technical assistance and support with cloning, expression and oxidative burst assays; Carlos Borroto, M.S., and Wan Hsin (Cindy) Chen, M.S. (Johns Hopkins University) for assistance with cloning; and Robert Cole, Ph.D. (Johns Hopkins University School of Medicine) A. B. Mass Spectrometry/Proteomic Facility for help with iTRAQ experiments and analyses.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00055/abstract

REFERENCES

Al-Khedery, B., Lundgren, A. M., Stuen, S., Granquist, E. G., Munderloh, U. G., Barbet, C. M., et al. (2012). Structure of the type IV secretion system in different strains of *Anaplasma phagocytophilum*. *BMC Genomics* 13:678. doi: 10.1186/1471-2164-13-678

Banerjee, R., Anguita, J., Roos, D., and Fikrig, E. (2000). Cutting edge: infection by the agent of human granulocytic ehrlichiosis prevents the respiratory burst by down-regulating gp91phox. *J. Immunol.* 164, 3946–3949. doi: 10.4049/jimmunol.164.8.3946

Beare, P. A., Howe, D., Cockrell, D. C., Omsland, A., Hansen, B., and Heinzen, R. A. (2009). Characterization of a *Coxiella burnetii ftsZ* mutant generated by Himar1 transposon mutagenesis. *J. Bacteriol.* 191, 1369–1381. doi: 10.1128/JB.01580-08

Borjesson, D. L., Kobayashi, S. D., Whitney, A. R., Voyich, J. M., Argue, C. M., and Deleo, F. R. (2005). Insights into pathogen immune evasion mechanisms: *Anaplasma phagocytophilum* fails to induce an apoptosis differentiation program in human neutrophils. *J. Immunol.* 174, 6364–6372. doi: 10.4049/jimmunol.174.10.6364

Carlyon, J. A., Chan, W. T., Galan, J., Roos, D., and Fikrig, E. (2002). Repression of *rac2* mRNA expression by *Anaplasma phagocytophila* is essential to the inhibition of superoxide production and bacterial proliferation. *J. Immunol.* 169, 7009–7018. doi: 10.4049/jimmunol.169.12.7009

Carlyon, J. A., and Fikrig, E. (2006). Mechanisms of evasion of neutrophil killing by *Anaplasma phagocytophilum*. *Curr. Opin. Hematol.* 13, 28–33. doi: 10.1097/01.moh.0000190109.00532.56

Caturegli, P., Asanovich, K. M., Walls, J. J., Bakken, J. S., Madigan, J. E., Popov, V. L., et al. (2000). *ankA*: an *Ehrlichia phagocytophila* group gene encoding a cytoplasmic protein antigen with ankyrin repeats. *Infect. Immun.* 68, 5277–5283. doi: 10.1128/IAI.68.9.5277-5283.2000

Chen, G., Severo, M. S., Sakhon, O. S., Choy, A., Herron, M. J. Pedra, F. R., et al. (2012). *Anaplasma phagocytophilum* dihydrolipoamide dehydrogenase 1 affects host-derived immunopathology during microbial colonization. *Infect. Immun.* 80, 3194–3205. doi: 10.1128/IAI.00532-12

Choi, K. S., Park, J. T., and Dumler, J. S. (2005). *Anaplasma phagocytophilum* delay of neutrophil apoptosis through the p38 mitogen-activated protein kinase signal pathway. *Infect. Immun.* 73, 8209–8218. doi: 10.1128/IAI.73.12.8209-8218.2005

Donnes, P., and Hoglund, A. (2004). Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics* 2, 209–215.

Dunning Hotopp, J. C., Lin, M., Madupu, R., Crabtree, J., Angiuoli, S. V., Eisen, J. A., et al. (2006). Comparative genomics of emerging human ehrlichiosis agents. *PLoS Genet.* 2:e21. doi: 10.1371/journal.pgen.0020021

Ellison, M. A., Thurman, G. W., and Ambruso, D. R. (2012). Phox activity of differentiated PLB-985 cells is enhanced, in an agonist specific manner, by the PLA2 activity of Prdx6-PLA2. *Eur. J. Immunol.* 42, 1609–1617. doi: 10.1002/eji.201142157

Felsheim, R. F., Herron, M. J., Nelson, C. M., Burkhardt, N. Y., Barbet, A. F., Kurtti, T. J., et al. (2006). Transformation of *Anaplasma phagocytophilum*. *BMC Biotechnol.* 6:42. doi: 10.1186/1472-6750-6-42

Garcia-Garcia, J. C., Barat, N. C., Trembley, S. J., and Dumler, J. S. (2009a). Epigenetic silencing of host cell defense genes enhances intracellular survival of the rickettsial pathogen *Anaplasma phagocytophilum*. *PLoS Pathog.* 5:488. doi: 10.1371/journal.ppat.1000488

Garcia-Garcia, J. C., Rennoll-Bankert, K. E., Pelly, S., Milstone, A. M., and Dumler, J. S. (2009b). Silencing of host cell *CYBB* gene expression by the nuclear effector AnkA of the intracellular pathogen *Anaplasma phagocytophilum*. *Infect. Immun.* 77, 2385–2391. doi: 10.1128/IAI.00023-09

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., et al. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 21, 617–623. doi: 10.1093/bioinformatics/bti057

Goodman, J. L., Nelson, C., Vitale, B., Madigan, J. E., Dumler, J. S., Kurtti, T. J., et al. (1996). Direct cultivation of the causative agent of human granulocytic ehrlichiosis. *N. Engl. J. Med.* 334, 209–215. doi: 10.1056/NEJM199601253340401

Hicks, G. R., and Raikhel, N. V. (1995). Protein import into the nucleus: an integrated view. *Ann. Rev. Cell. Dev. Biol.* 11, 155–188. doi: 10.1146/annurev.cb.11.110195.001103

Hoglund, A., Donnes, P., Blum, T., Adolph, H. W., and Kohlbacher, O. (2006). MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22, 1158–1165. doi: 10.1093/bioinformatics/btl002

IJdo, J. W., Carlson, A. C., and Kennedy, E. L. (2007). *Anaplasma phagocytophilum* AnkA is tyrosine-phosphorylated at EPIYA motifs and recruits SHP-1 during early infection. *Cell. Microbiol.* 9, 1284–1296. doi: 10.1111/j.1462-5822.2006.00871.x

Larson, C. L., Beare, P. A., Howe, D., and Heinzen, R. A. (2013). *Coxiella burnetii* effector protein subverts clathrin-mediated vesicular trafficking for pathogen vacuole biogenesis. *Proc. Natl. Acad. Sci. U*.S.A. 110, E4770–E4779. doi: 10.1073/pnas.1309195110

Lin, M., den Dulk-Ras, A., Hooykaas, P. J., and Rikihisa, Y. (2007). *Anaplasma phagocytophilum* AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection. *Cell. Microbiol.* 9, 2644–2657. doi: 10.1111/j.1462-5822.2007.00985.x

Lin, M., and Rikihisa, Y. (2003). *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum* lack genes for lipid A biosynthesis and incorporate cholesterol for their survival. *Infect. Immun.* 71, 5324–5331. doi: 10.1128/IAI.71.9.5324-5331.2003

Lodes, M. J., Mohamath, R., Reynolds, L. D., McNeill, P., Kolbert, C. P., Bruinsma, E. S., et al. (2001). Serodiagnosis of human granulocytic ehrlichiosis by using novel combinations of immunoreactive recombinant proteins. *J. Clin. Microbiol.* 39, 2466–2476. doi: 10.1128/JCM.39.7.2466-2476.2001

Mastronunzio, J. E., Kurscheid, S., and Fikrig, E. (2012). Postgenomic analyses reveal Development of infectious *Anaplasma phagocytophilum* during transmission from ticks to mice. *J Bacteriol.* 194, 2238–2247. doi: 10.1128/JB.06791-11

Nair, R., Carter, P., and Rost, B. (2003). NLSdb: database of nuclear localization signals. *Nucleic. Acids. Res.* 31, 397–399. doi: 10.1093/nar/gkg001

Nelson, C. M., Herron, M. J., Felsheim, R. F., Schloeder, B. R., Grindle, S. M., Chavez, A. O., et al. (2008). Whole geonme transcription profiling of *Anaplasma phagocytophilum* in human and tick host cells by tiling array analysis. *BMC Genomics* 9:364. doi: 10.1186/1471-2164-9-364

Ohashi, N., Zhi, N., Lin, Q., and Rikihisa, Y. (2002). Characterization and transcriptional analysis of gene clusters for a type IV secretion machinery in human granulocytic and monocytic ehrlichiosis agents. *Infect. Immun.* 70, 2128–2138. doi: 10.1128/IAI.70.4.2128-2138.2002

Park, J., Kim, K. J., Choi, K. S., Grab, D. J., and Dumler, J. S. (2004). *Anaplasma phagocytophilum* AnkA binds to granulocyte DNA and nuclear proteins. *Cell. Microbiol.* 6, 743–751. doi: 10.1111/j.1462-5822.2004.00400.x

Pedruzzi, E., Fay, M., Elbim, C., Gaudry, M., and Gougerot-Pocidalo, M. A. (2002). Differentiation of PLB-985 myeloid cells into mature neutrophils, shown by degranulation of terminally differentiated compartments in response to N-formyl peptide and priming of superoxide anion production by granulocyte-macrophage colony-stimulating factor. *Br. J. Haematol.* 117, 719–726. doi: 10.1046/j.1365-2141.2002.03521.x

Rennoll-Bankert, K. E., and Dumler, J. S. (2012). Lessons from *Anaplasma phagocytophilum*: chromatin remodeling by bacterial effectors. *Infect. Dis. Drug. Targets* 12, 380–387. doi: 10.2174/187152612804142242

Rennoll-Bankert, K. E., Sinclair, S. H., Lichay, M. A., and Dumler, J. S. (2014). Comparison and characterization of granulocyte cell models for *Anaplasma phagocytophilum* infection. *Pathog. Dis.* 71, 55–64. doi: 10.1111/2049-632X.12111

Rikihisa, Y., Lin, M., and Niu, H. (2010). Type IV secretion in the obligatory intracellular bacterium *Anaplasma phagocytophilum*. *Cell. Microbiol.* 12, 1213–1221. doi: 10.1111/j.1462-5822.2010.01500.x

Sinclair, S. H., Rennoll-Bankert, K. E., and Dumler, J. S. (2014). Effector bottleneck: microbial reprogramming of parasitized host cell transcription by epigenetic remodeling of chromatin structure. *Front. Genet.* 5:274. doi: 10.3389/fgene.2014.00274

Troese, M. J., Kahlon, A., Ragland, S. A., Ottens, A. K., Ojogun, N., Carlyon, J. A., et al. (2011). Proteomic analysis of *Anaplasma phagocytophilum* during infection of human myeloid cells identifies a protein that is pronouncedly upregulated on the infectious dense-cored cell. *Infect. Immun.* 79, 4696–4707. doi: 10.1128/IAI.05658-11

Voth, D. E., Beare, P. A., Howe, D., Sharma, U. M., Samoilis, G., Cockrell, D. C., et al. (2011). The *Coxiella burnetii* cryptic plasmid is enriched in genes encoding type IV secretion system substrates. *J. Bacteriol.* 193, 1493–1503. doi: 10.1128/JB.01359-10

Wang, X., Cheng, Z., Zhang, C., Kikuchi, T., and Rikihisa, Y. (2007). *Anaplasma phagocytophilum p44* mRNA expression is differentially regulated in mammalian and tick host cells: involvement of the dna binding protein ApxR. *J. Bacteriol.* 189, 8651–8659. doi: 10.1128/JB.00881-07

Yang, F., Vought, B. W., Satterlee, J. S., Walker, A. K., Jim Sun, Z. Y., Watts, J. L., et al. (2006). An ARC/Mediator subunit required for SREBP control of cholesterol and lipid homeostasis. *Nature* 442, 700–704. doi: 10.1038/nature04942

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 December 2014; accepted: 16 January 2015; published online: 06 February 2015.

*Citation: Sinclair SHG, Garcia-Garcia JC and Dumler JS (2015) Bioinformatic and mass spectrometry identification of Anaplasma phagocytophilum proteins translocated into host cell nuclei. Front. Microbiol. **6**:55. doi: 10.3389/fmicb.2015.00055*

This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 June 2015
doi: 10.3389/fmicb.2015.00608

Bottom-up modeling approach for the quantitative estimation of parameters in pathogen-host interactions

Teresa Lehnert[1,2†], *Sandra Timme*[1,2†], *Johannes Pollmächer*[1,2], *Kerstin Hünniger*[3], *Oliver Kurzai*[2,3] *and Marc Thilo Figge*[1,2*]

[1] *Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute, Jena, Germany,* [2] *Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, Jena, Germany,* [3] *Fungal Septomics, Septomics Research Center, Friedrich Schiller University and Leibniz Institute for Natural Product Research and Infection Biology Hans-Knöll-Institute, Jena, Germany*

OPEN ACCESS

Edited by:
Saliha Durmus, Gebze Technical University, Turkey

Reviewed by:
Reiko Tanaka, Imperial College London, UK
Muhammed Erkan Karabekmez, Bogazici University, Turkey

***Correspondence:**
Marc Thilo Figge, Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Adolf-Reichwein-Straße 23, Beutenberg Str 11a, 07745 Jena Germany
thilo.figge@hki-jena.de

[†] *These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *15 April 2015*
Accepted: *02 June 2015*
Published: *19 June 2015*

Citation:
Lehnert T, Timme S, Pollmächer J, Hünniger K, Kurzai O and Figge MT (2015) Bottom-up modeling approach for the quantitative estimation of parameters in pathogen-host interactions. Front. Microbiol. 6:608. doi: 10.3389/fmicb.2015.00608

Opportunistic fungal pathogens can cause bloodstream infection and severe sepsis upon entering the blood stream of the host. The early immune response in human blood comprises the elimination of pathogens by antimicrobial peptides and innate immune cells, such as neutrophils or monocytes. Mathematical modeling is a predictive method to examine these complex processes and to quantify the dynamics of pathogen-host interactions. Since model parameters are often not directly accessible from experiment, their estimation is required by calibrating model predictions with experimental data. Depending on the complexity of the mathematical model, parameter estimation can be associated with excessively high computational costs in terms of run time and memory. We apply a strategy for reliable parameter estimation where different modeling approaches with increasing complexity are used that build on one another. This bottom-up modeling approach is applied to an experimental human whole-blood infection assay for *Candida albicans*. Aiming for the quantification of the relative impact of different routes of the immune response against this human-pathogenic fungus, we start from a non-spatial state-based model (SBM), because this level of model complexity allows estimating *a priori* unknown transition rates between various system states by the global optimization method *simulated annealing*. Building on the non-spatial SBM, an agent-based model (ABM) is implemented that incorporates the migration of interacting cells in three-dimensional space. The ABM takes advantage of estimated parameters from the non-spatial SBM, leading to a decreased dimensionality of the parameter space. This space can be scanned using a local optimization approach, i.e., *least-squares error estimation* based on an *adaptive regular grid search*, to predict cell migration parameters that are not accessible in experiment. In the future, spatio-temporal simulations of whole-blood samples may enable timely stratification of sepsis patients by distinguishing hyper-inflammatory from paralytic phases in immune dysregulation.

Keywords: state-based model, agent-based model, pathogen-host interaction, parameter estimation, whole-blood infection assay, *Candida albicans*

1. Introduction

The human fungal pathogen *Candida albicans* is part of the normal microbial flora in more than half of the global population. In immunocompromised patients it can become invasive and may enter the blood stream via medical devices, e.g., catheters, or translocation in the gut and can cause severe systemic infections. The immune response against *C. albicans* in human blood comprises the interplay of various complex biological processes involving different immune mechanisms (Duggan et al., 2015b). Most importantly, the whole-blood infection assay allows multiple immune effector mechanisms to occur at the same time and thus modulate the overall outcome (Luo et al., 2013; Cunha et al., 2014; Hünniger et al., 2015). Applying a systems biology approach, we quantified individual processes and in this way revealed the main route of the immune response against *C. albicans* in human blood (Hünniger et al., 2014). This was achieved by an iterative systems biology cycle involving experiment, mathematical modeling, hypothesis generation and further experimental investigation.

The choice of an appropriate mathematical modeling approach strongly depends on the questions to be answered and the hypothesis, as well as the characteristics of the underlying experimental data with regard to temporal and spatial information. A wide range of modeling approaches exists that differ by their computational complexity and can be classified depending on the degree of spatial representation as well as the internal degrees of freedom attributed to the model entities. The computationally cheapest modeling approach for dynamic systems is represented by ordinary differential equations (ODE), where biological entities are assumed to be present in high numbers and spatial information is not required such that they can be collectively represented by a homogeneously distributed concentration variable. State-based models (SBM) resolve the biological entities as individuals that occupy states and are able to perform transitions between states representing dynamic processes. In contrast to ODE, this approach allows modeling discrete events for any entity number in a biological system. However, SBM are in turn limited in that they do not represent spatial aspects. Individual-based models (IBM) such as cellular automata (CA) and agent-based models (ABM) do simulate discrete entities in space and time (Medyukhina et al., 2015). In a CA simulation, these entities can undergo state changes associated with their internal degrees of freedom as well as positional changes on a pre-defined spatial grid of computational cells (Von Neumann, 1951; Bittig and Uhrmacher, 2010). The discrete number of individual entities as well as the spatial representation of the environment result in increasing computational costs in terms of run-time and memory. Even more computationally expensive but biologically more realistic simulations can be performed by the ABM approach. Here, biological objects are represented as individual entities, so-called agents, that are able to move in space and can act as well as interact with other agents according to individual properties. Examples of ABM for the pathogen-host interaction between the human-pathogenic fungus *Aspergillus fumigatus* and phagocytes were presented by Tokarski et al. (2012) and Pollmächer and Figge (2014). In particular, the ABM developed by Pollmächer and Figge (2014) simulates the detection of *A. fumigatus* conidia by macrophages in a to-scale representation of human alveoli and predicts the requirement of a chemotactic signal guiding the phagocytes to the spatial positions of conidia.

In general, parameters of bio-mathematical models characterize the components by their morphology and their dynamic behavior. For example, cells may be defined by parameters for size and shape as well as by parameters for interactions in the spatial environment that are associated with the typical frequency of interaction processes. Model parameters associated with dynamical, functional and morphological aspects of biological processes may be extracted from microscopic images by applying an image-based systems biology approach (Horn et al., 2012; Mech et al., 2014; Medyukhina et al., 2015). However, in many cases microscopy experiments cannot be performed for technical reasons, as is also the case for whole-blood infection assays where the majority of cells are erythrocytes blocking the view on leukocytes, let alone fungal pathogens that are present in even lower numbers. In situations like these, numerical estimation of *a priori* unknown parameter values by comparison with experimental time-series data becomes a highly relevant issue. Parameter estimation algorithms are applied to find the optimal match between the experimental data and simulated model data. These optimization algorithms can be characterized by their search technique within the parameter space, i.e., as global or local approaches, and their mathematical procedures, i.e., as stochastic or deterministic approaches (Moles et al., 2003; Ashyraliyev et al., 2009). Local optimization techniques search for better parameter values within a locally restricted parameter space, where the direct search method and gradient based methods are widely used (Ashyraliyev et al., 2009). They show fast convergence to the optimal parameter values, but since local optimization algorithms will get stuck in a nearby local optimum, an educated guess of the initial parameter values is absolutely required. In contrast, global optimization strategies search a wide range of the parameter space with possibly various local optima and the subclass of deterministic optimization strategies can find the global optimum with pre-defined accuracy (Ashyraliyev et al., 2009). High-dimensional parameter spaces may be searched by stochastic optimization algorithms that make use of probabilistic elements to avoid getting trapped in local optima in order to find the global optimum. Common stochastic search algorithms of this type are Metropolis Monte Carlo (MMC) (Metropolis et al., 1953), adaptive random search and evolutionary computation techniques such as differential evolution (DE) (Storn and Price, 1997). Additionally, heuristics can be applied in support of a fast convergence rate of global or local optimization strategies, e.g., simulated annealing (SA) (Kirkpatrick et al., 1983; Gonzalez et al., 2007), great deluge (Dueck, 1993), or performing multiple searches from random start parameters. The selection of the most suitable optimization algorithm depends on specific model properties, such as the dimension of the parameter space and the computational costs for the model simulations that have to be repeatedly performed. For computationally cheap ODE models, the computationally expensive stochastic global optimization algorithms may be used,

such as DE applied by Hernandez-Vargas et al. (2014) and SA based on MMC applied by Hünniger et al. (2014) and Mech et al. (2014).

The non-spatial virtual infection model of the immune response against *C. albicans* in human blood was formulated as a SBM and its parameters were fitted to the experimentally determined time-evolution of concentrations for *C. albicans* cells that are alive or killed and that can either reside in extracellular space or inside immune cells of different types, i.e., monocytes or granulocytes (polymorphonuclear neutrophils, PMN) (Hünniger et al., 2014). Furthermore, we observed a cell population of *C. albicans* that remained alive or killed in extracellular space, i.e., these fungal cells are resistant against phagocytosis and/or killing. The different *C. albicans* cell populations were assigned states and individual cells could perform transitions between states, such as phagocytosis by immune cells, subsequent intracellular killing, extracellular killing by antimicrobial peptides or acquiring resistance against phagocytosis and/or killing. Resistant *C. albicans* cells are a population of cells that were found to be protected against phagocytosis and/or killing and that remained in the extracellular space of the whole-blood infection assay (Hünniger et al., 2014). Since the model is restricted to the dynamics of states occupied by pathogenic cells we refer to the model by Hünniger et al. (2014) as P-SBM. In the present study, motivated by newly measured experimental data regarding the immune cell number of monocytes and PMN in the whole-blood assays, we take the next step and modify the P-SBM to drop its implicit assumption that the number of immune cells for samples from different individuals would be the same. Since in the modified SBM states are assigned to the pathogenic cells as well as to the two types of immune cells, which have been found to actively participate in *C. albicans* elimination, we will refer to this model as PI-SBM. Taking individual immune cells explicitly into account obviously makes the simulations of the whole-blood infection assay more realistic, albeit at the expense of higher computational costs for global parameter optimization that will be performed using SA based on the MMC scheme as was the case for the P-SBM.

A timely stratification of sepsis patients in different phases of immune dysregulation requires spatio-temporal simulations of whole-blood samples. To achieve this goal, an ABM of the whole-blood infection assay was established that builds on the PI-SBM and incorporates spatial properties of the blood sample in a three-dimensional continuous representation. In particular, in the ABM *C. albicans* cells as well as monocytes and PMN are agents that can migrate in the environment and interact with each other. Apart from the model parameters associated with the migration of cells, the ABM was based on the transition rates of the PI-SBM after appropriate conversion. This procedure strongly reduces the number of *a priori* unknown parameters of agents to the subset of migration parameters. The latter can be estimated using the computationally cheap grid search algorithm and enables the prediction of the migration behavior for the different immune cell types that are otherwise not directly accessible in experiment. The interrelations between the different modeling approaches are schematically shown in **Figure 1** demonstrating that results are re-used across different modeling approaches to simultaneously facilitate an increase in model complexity and a decrease in computational expense for parameter estimation. Our step-wise computational biology approach avoids typical limitations of realistic models by focusing parameter estimation on those parameters that arise at the next level of model complexity.

2. Materials and Methods

2.1. Non-spatial State-based Model

The initial version of the non-spatial SBM describes the dynamics of state transitions for the human-pathogenic fungus *C. albicans* in whole-blood samples of healthy donors (Hünniger et al., 2014). In agreement with experimental data, the time-evolution of different *C. albicans* cells that are alive or killed and in extracellular space or phagocytosed by either monocytes or PMN can be simulated in this way. Since this SBM assumes the number of immune cells to be constant across blood samples of different donors and does only simulate the dynamics of the pathogenic (P) cells, it is hereafter referred to as P-SBM. However, it is known that the number of immune cells may strongly vary across human individuals and in particular for patients. Therefore, we increase the model complexity by advancing the P-SBM to a model that does explicitly account for the number of immune cells being present in a hemogram. Data including immune cell counts can easily be obtained both in an experimental as well as in a clinical setting. This model is hereafter referred to as PI-SBM to indicate that state transitions are computed for pathogenic (P) as well as immune (I) cells.

For comparison between the model predictions and the experimentally determined kinetics in the whole-blood infection assay, we introduce specific combinations of states, referred to as *combined units*, that are measurable and useable for the parameter estimation. These comprise all extracellular *C. albicans* cells C_E,

$$C_E \equiv C_{AE} + C_{KE} + C_{AR} + C_{KR}\,, \tag{1}$$

that are either alive (C_{AE}) or killed (C_{KE}) cells in extracellular space as well as cells resistant against killing and/or phagocytosis that are either alive (C_{AR}) or killed (C_{KR}). Next, the combined units C_M and C_G refer to *C. albicans* cells that are phagocytosed, respectively, by monocytes

$$C_M \equiv \sum_{i\geq 0}\sum_{j\geq 0} M_{i,j}\,(i+j)\,, \tag{2}$$

or by granulocytes

$$C_G \equiv \sum_{i\geq 0}\sum_{j\geq 0} G_{i,j}\,(i+j)\,. \tag{3}$$

Here, $M_{i,j}$ and $G_{i,j}$ refer to the number of monocytes and granulocytes (PMN), respectively, with i alive and j killed phagocytosed *C. albicans* cells. We limit the maximal number of *C. albicans* cells that can be phagocytosed by an immune

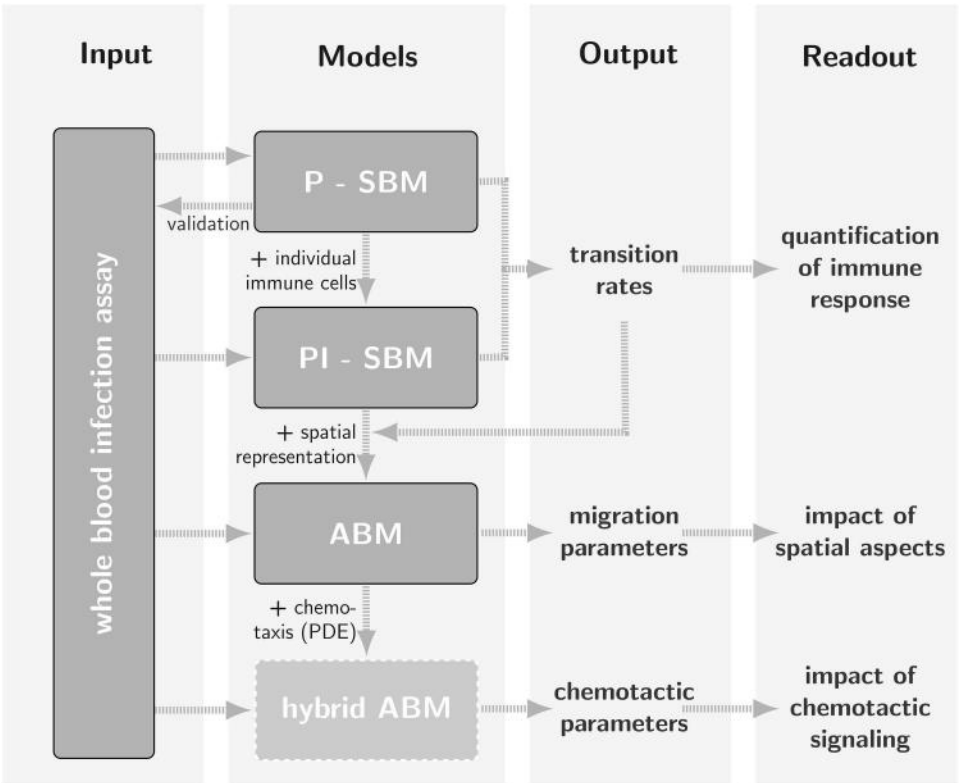


FIGURE 1 | Schematic representation of the bottom-up modeling approach as a strategy for parameter estimation where different mathematical models with increasing levels of complexity build on one another. Using model outputs at one level for calibration of models at a higher level is by way of example demonstrated for *Candida albicans* infection in a human whole-blood assay (Hünniger et al., 2014). The state-based model P-SBM, which focuses on the viewpoint of pathogens, was modified into the state-based model PI-SBM by accounting for actions of individual immune cells. After calibration of transition rates in the non-spatial state-based model, these were used to simulate the infection process in an agent-based model (ABM) that accounts for the spatial representation of the whole-blood infection assay. In the future, the calibration of migration parameters may for example serve as input for a hybrid ABM that captures the time-evolution of chemotactic signaling by solving reaction-diffusion equations.

cell to 18, i.e., $i, j < 10$, being much larger than observed in experiment (Hünniger et al., 2014). Furthermore, all killed *C. albicans* cells are given by the combined unit

$$C_K \equiv C_{KE} + C_{KR} + \sum_{i\geq 0}\sum_{j\geq 1}(M_{i,j} + G_{i,j})\, j\,, \tag{4}$$

and all alive *C. albicans* cells by the combined unit

$$C_A \equiv C_{AE} + C_{AR} + \sum_{i\geq 1}\sum_{j\geq 0}(M_{i,j} + G_{i,j})\, i\,. \tag{5}$$

It should be noted that only three of the five combined units are independent of each other, due to the conservation relations $C = C_E + C_G + C_M$ and $C = C_K + C_A$ for the total number of *C. albicans* cells C.

The simulation algorithm for the time-evolution of the PI-SBM is implemented in C++ that is available upon request. In **Figure 2A**, the simulation algorithm is schematically depicted and can be compared to the simulation algorithm of the P-SBM in Supplementary Figure 1. We simulate a blood sample of 1 ml containing 5×10^5 monocytes, 5×10^6 PMN and 1×10^6 *C. albicans* cells that are initially extracellular and alive. In each time-step, which we set to $\Delta t_{PI-SBM} = 1$ min, the algorithm tests for each individual cell in the system whether or not it does undergo a state transition. To this end, a cell is first randomly selected by sampling its relative frequency of occurrence among all cell types in the system. Next, the state of this cell is updated using a random selection procedure for the one transition in this time-step that the cell can possibly make among all currently enabled transitions. Once the type of transition between states s and s' with rate $r_{s\to s'}$ is selected, it will be executed with probability $P_{s\to s'} = r_{s\to s'}\,\Delta t_{PI-SBM}$ and the system is updated accordingly. **Table 1** provides an overview of the transition rates for all possible state transitions of the model. After testing all individuals in the system for performing a state transition, the simulation time is advanced by one time-step and the whole procedure is repeated until the total simulation time is reached. Note that, since the ratio of the number of immune cells over the number of pathogenic cells is larger than five, the simulation run time of the PI-SBM is significantly increased compared with the P-SBM.

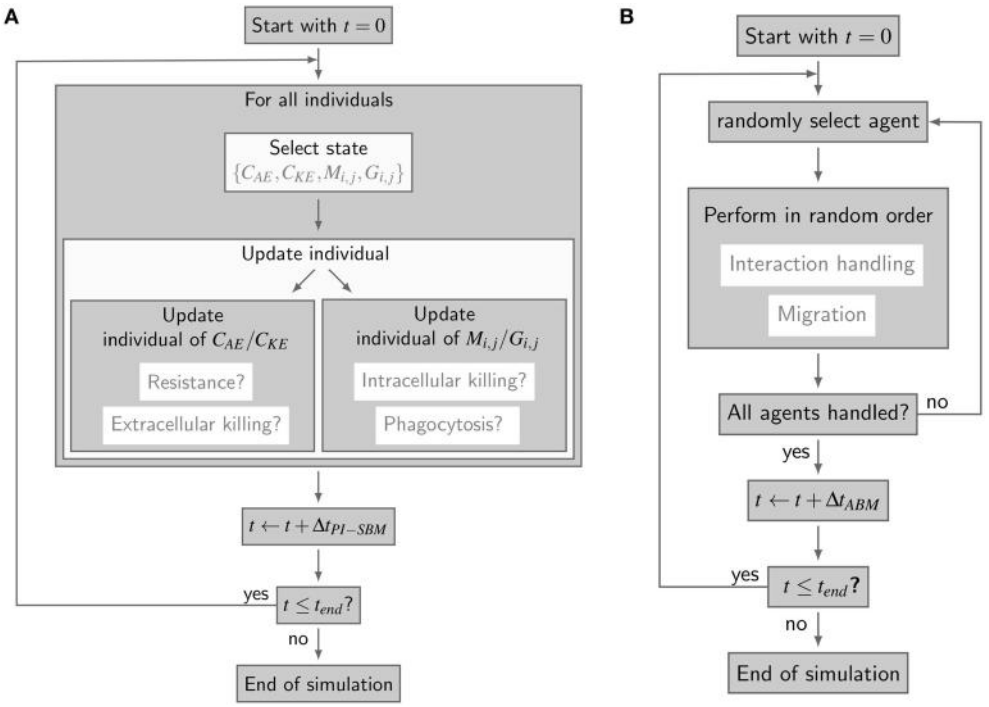


FIGURE 2 | Simulation algorithms of virtual infection models for whole-blood assays. (A) Flow-chart of the non-spatial PI-SBM simulation algorithm. In each time-step Δt_{PI-SBM}, all individuals are tested for possible state transitions. Individuals of extracellular alive and killed *C. albicans* states, i.e., C_{AE} and C_{KE}, respectively, are tested for becoming resistant and for extracellular killing. Individuals of immune cell states ($M_{i,j}$ or $G_{i,j}$) are tested for phagocytosis of *C. albicans* and for intracellular killing. **(B)** Flow chart of the spatial ABM simulation algorithm. In each time-step Δt_{ABM}, the migration and interaction handling is performed in random order for every randomly chosen agent.

2.2. Spatial Agent-based Model

The spatial virtual infection model for *C. albicans* in human blood is realized using an ABM approach. This model is implemented in C++ based on a previously established framework of Pollmächer and Figge (2014) and is the spatial counterpart of the non-spatial PI-SBM introduced in Section 2.1. The C++ source code of the ABM simulation algorithm is available upon request. In the ABM, the two types of immune cells—monocytes and PMN—as well as the pathogenic *C. albicans* cells are incorporated as virtual objects. These virtual objects are agents that are characterized by a spherical morphology with the physiological diameters of $d_M = 16\,\mu$m for monocytes, $d_G = 13.5\,\mu$m for PMN (Mak and Saunders, 2011) and $d_C = 7\,\mu$m for *C. albicans* (Mendling, 2006) (see **Figure 3A**) and that can migrate and interact with each other on encounter in the three-dimensional spatial environment (see **Figure 3B**). We impose a cuboid environment with an edge length of $1000\,\mu$m representing $1\,\mu$l blood and use *random periodic* boundary conditions for the cuboid, i.e., an agent which leaves the environment at some boundary point is deleted from the system and a new agent with identical properties re-enters the environment at some other randomly chosen boundary point. The cuboid environment is represented as a continuous space, i.e., allowing agents to move in a manner that is more realistic than could be captured by a lattice-based approach. This advantage is accompanied by the drawback that well-defined neighborhood relations as naturally existing between neighboring sites on a lattice are not present in continuous space representations. However, in order to efficiently determine cell–cell encounters, we use a neighborhood list method, which reduces the computational complexity to a close-to linear dependency on the number of agents in the system (Rapaport, 1996). At time point $t = 0$, agents are initialized with all *C. albicans* cells being in the state alive-and-extracellular. The time-evolution of the system is simulated by the random selection method (Skvoretz, 2002; Figge, 2005) that handles the migration and interaction of agents per time-step Δt in a random fashion (see **Figure 2B**).

We use ratios in cell numbers that are equivalent to those in the PI-SBM, where $1\,\mu$l of blood contains 5×10^3 PMN, 5×10^2 monocytes and 1×10^3 *C. albicans* cells, i.e., in total 6.5×10^3 cells. Viewing cells as interacting point particles, an average volume of $v \approx \frac{1}{6.5}\times10^6\,\mu\text{m}^3$ can be attributed to each cell, implying an average distance of $l \approx v^{1/3} \approx 55\,\mu$m between immune cells and *C. albicans* cells. Even though this distance is clearly larger than the diameters of these cells, $l \gg d_M, d_G, d_C$, we assume that the migration behavior of immune cells and *C. albicans* cells in blood resembles a random walk of agents without directional persistence. This assumption is based on the fact that the total number of erythrocytes in human blood ranges from $4\times10^6 - 6\times10^6$ cells/μl (McClatchey, 2003). Estimating the total number of cells in $1\,\mu$l of blood to be about six millions, an average volume

TABLE 1 | Rates of state transitions in the non-spatial PI-SBM.

Transition rate	Description	State transition
ϕ_M	Phagocytosis by monocytes	$M_{i,j} + C_{AE} \to M_{i+1,j}$ $M_{i,j} + C_{KE} \to M_{i,j+1}$
κ_M	Intracellular killing by monocytes	$M_{i,j} \to M_{i-1,j+1}$
ϕ_G	Phagocytosis by PMN for first-time phagocytosis event	$G_{0,0} + C_{AE} \to G_{1,0}$ $G_{0,0} + C_{KE} \to G_{0,1}$
$\phi_{G^\star}$	Phagocytosis by PMN for repeated phagocytosis events	$G_{i,j} + C_{AE} \to G_{i+1,j}$ $G_{i,j} + C_{KE} \to G_{i,j+1}$
κ_G	Intracellular killing by PMN	$G_{i,j} \to G_{i-1,j+1}$
$\kappa_{EK}(t)$	Extracellular killing by antimicrobial peptides released by first-time PMN phagocytosis with decreasing activity Rate depends on the activity of antimicrobial peptides ($\bar{\kappa}_{EK}$) and the decay of their antimicrobial activity (γ) as defined in Hünniger et al. (2014)	$C_{AE} \to C_{KE}$
ρ	Resistance against phagocytosis and/or killing	$C_{AE} \to C_{AR}$, $C_{KE} \to C_{KR}$

For details see (Hünniger et al., 2014).

of $v_c \approx \frac{1}{6}\times 10^3\,\mu\text{m}^3$ can be attributed to each cell, implying a mean free path of $l_{fp} \approx v_c^{1/3} \approx 5\,\mu$m between point particles. This distance is not only clearly smaller than the distance between immune cells and *C. albicans* cells, $l_{fp} \ll l$, but also smaller than the diameters of erythrocytes, *C. albicans* cells as well as of the immune cells under consideration. It can be concluded that cells are not migrating with directional persistence in blood, because frequent collisions with the overwhelming number of erythrocytes will induce diffusive migration of cells with diffusion coefficients in whole-blood that can be very different for the different cell types. This is a consequence of the fact that monocytes and PMN perform active migration, whereas *C. albicans* cells are immotile due to the complete lack of cellular organelles for motility (Margulies and Schwartz, 1998) and its movement in whole blood is only passive.

Even though blood is a non-Newtonian fluid, i.e., showing pseudoplastic properties with variable viscosity depending on the exerted shear stress in capillaries of different sizes (Fahraus and Lindqvist, 1931), the experimental setup of the whole-blood infection assay is such that the viscosity as well as the temperature in the mildly stirred test tube remain constant (Hünniger et al., 2014). Therefore, the Stokes-Einstein equation (Einstein, 1905) can be applied to infer the diffusion coefficient D_C for the passive movement of *C. albicans* cells. Based on a whole-blood viscosity of about $\eta \approx 4\,mPa\,s$ (Rosenson et al., 1996), Boltzmann constant k_B and temperature $T = 37°$C (Hünniger et al., 2014), this yields the relatively small diffusion coefficient $D_C = k_B T/(3\pi\eta d_C) \approx 1\,\mu\text{m}^2/\text{min}$. In contrast, the active migration of monocytes and PMN requires to estimate their diffusion coefficients numerically.

The time-step Δt_{ABM} for simulations in the ABM has to be chosen such that a smooth migration of cells is sampled in time. In order to ensure this, we require that during one time-step Δt_{ABM} cells do not migrate further than a certain distance, which we set to equal the mean free path $l_{fp} = 5\,\mu$m:

$$\Delta t_{ABM} = \frac{l_{fp}^2}{6\,D_{max}}\,. \tag{6}$$

Here, $D_{max} \equiv \max\{D_C, D_M, D_G\}$ denotes the largest out of the three diffusion coefficients for *C. albicans* cells (D_C), monocytes (D_M), and PMN (D_G). Since it can be expected that the active migration of immune cells is associated with diffusion coefficients D_M and D_G with $D_M, D_G \gg 1\,\mu\text{m}^2/\text{min}$ in the whole-blood infection assay, it follows from Equation (6) that the time-step in the ABM will be much smaller than in the state-based model PI-SBM: $\Delta t_{ABM} \ll \Delta t_{PI-SBM} = 1$ min. Moreover, stochasticity in the ABM requires that each simulation has to be repeated multiple times, resulting into relatively high computational costs compared with the PI-SBM, in particular, if we would have envisaged to estimate each model parameter instead of following the strategy of a bottom-up modeling approach.

Computational costs associated with parameter estimation in the ABM can be significantly reduced by making use of the previously estimated rates of state transitions in the state-based model PI-SBM (see Section 2.1 and **Table 1**). In the course of a simulation, migrating cells in the ABM may either spontaneously undergo state transitions or interact with each other upon spatial contact. In **Figure 3C**, we present a schematic overview of processes that occur according to defined rules associated with certain probabilities. It is important to note that, due to the spatial aspects that are captured by the ABM but not the PI-SBM, we have to distinguish between processes that are *contact-dependent* and *contact-independent*.

For *contact-independent* processes—such as intracellular and extracellular killing as well as the occurrence of *C. albicans* resistance against phagocytosis and/or killing—the conversion of rates from the PI-SBM to the ABM is straightforward. Since these processes are not determined by any spatial requirements, a simple re-scaling is performed. For example, *C. albicans* cells become resistant in the PI-SBM with probability $P_{PI-SBM}(\rho) = \rho\,\Delta t_{PI-SBM}$. In the ABM, where the resolution of time is set by the time-step $\Delta t_{ABM} \ll \Delta t_{PI-SBM}$, we check in each time-step with probability

$$P_{ABM}(\rho) = P_{PI-SBM}(\rho)\frac{\Delta t_{ABM}}{\Delta t_{PI-SBM}} \tag{7}$$

whether this process occurs.

In contrast, *contact-dependent* processes in the ABM are characterized by the requirement that two cells have to get into spatial contact first, before such a process—for example, a phagocytosis event of a *C. albicans* cell by a monocyte with transition rate ϕ_M—can take place. In the PI-SBM, spatial contact is not explicitly modeled; rather, the interaction partner for each monocyte is randomly chosen once per time-step Δt_{PI-SBM}. The associated probability is determined by the time-dependent ratio of non-resistant fungal cells over the sum of extracellular fungal cells and immune cells. Once an interaction partner was chosen, the phagocytosis event itself occurs with probability

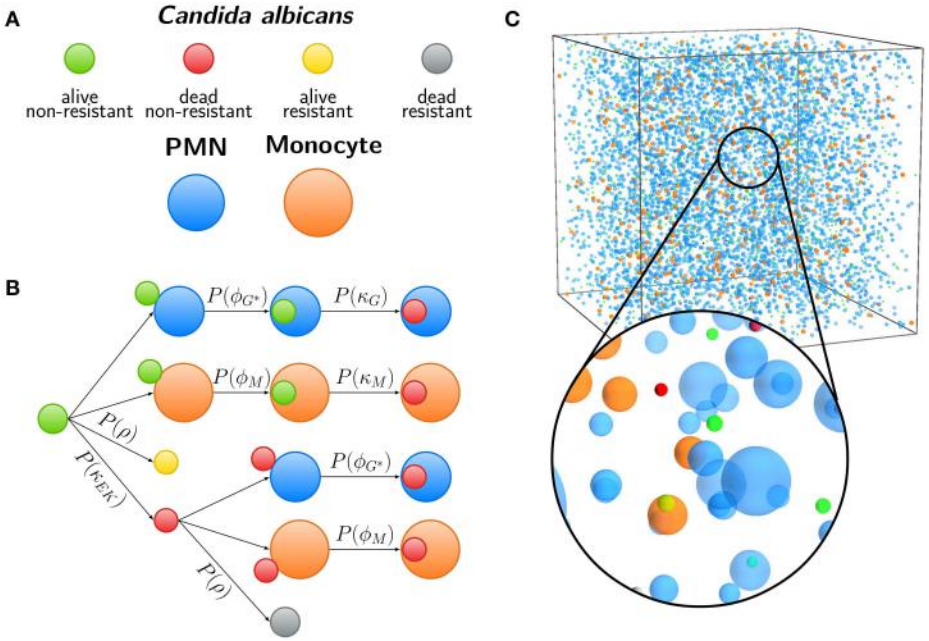


FIGURE 3 | Agent-based model (ABM) of the human whole-blood infection assay. (A) Agents used in the ABM are *Candida albicans* cells in the four states (i) alive and non-resistant (green), (ii) dead and non-resistant (red), (iii) alive and resistant (yellow), and (iv) dead and resistant (gray). Furthermore, granulocytes (polymorphonuclear neutrophils, PMN) and monocytes are shown in blue and orange, respectively. **(B)** Schematic overview of examples for interactions in the ABM. Each arrow indicates the execution of an interaction event, either depending on spatial contact between different cells or by a contact-independent process. *C. albicans* cells that are in extracellular space and alive (C_{AE}) or killed (C_{KE}) can become resistant with probability $P(\rho)$. After spatial contact, alive or killed *C. albicans* cells can be phagocytosed by PMN or monocytes with probabilities $P(\phi_G|\phi_{G^*})$ or $P(\phi_M)$, respectively. Intracellular alive *C. albicans* cells are killed with probabilities $P(\kappa_G)$ or $P(\kappa_M)$ depending on the type of phagocyte. **(C)** Visualization of the three-dimensional cuboid environment of the ABM that corresponds to 1 μl of the whole-blood infection assay, containing 5000 PMN, 500 monocytes, and 1000 *C. albicans* cells. The time-evolution of the simulated infection scenario can be viewed in Supplementary Video 1.

$P_{PI-SBM}(\phi_M) = \phi_M \, \Delta t_{PI-SBM}$ in the PI-SBM. Correspondingly, in the ABM, we request that this process takes place with the same probability,

$$P_{ABM}(\phi_M) = P_{PI-SBM}(\phi_M)\,, \tag{8}$$

on every encounter between a monocyte and a *C. albicans* cell. This correspondence of event probabilities for the two modeling approaches imposes a condition on the spatial dynamics of cells, i.e., on the values of the diffusion coefficients in the ABM and by that on the time-step Δt_{ABM} (see Equation 6). For optimal migration parameters, i.e., parameters that result in good agreement with the experimental data, it is expected that measurement of the associated phagocytosis rate in the ABM coincides with the corresponding rate from the PI-SBM.

2.3. Parameter Estimation

2.3.1. Simulated Annealing

The *a priori* unknown transition rates of the PI-SBM are determined by the method of Simulated Annealing based on the Metropolis Monte Carlo scheme (SA-MMC) that is depicted in **Figure 4A**. This optimization method randomly explores the parameter space of transition rates to find the global minimum of the fitting error, i.e., the most suitable parameter set that produces the best fit of the simulation to the experimental data obtained from the whole-blood infection assay.

The parameter estimation algorithm starts with a randomly chosen set of parameter values within the interval of $[0, 1]$ per minute, represented by the vector $\vec{p}$, and calculates the resulting time-evolution of state occupations from the simulation algorithm of the PI-SBM (see Section 2.1). To score the simulation result for a particular set of parameters, we combined different kinetics of the PI-SBM, referred to as *combined units*, which are identical with the experimental kinetics measured in the whole-blood infection assay (see Section 2.1). In this way, the experimental kinetics can be directly compared with the combined units c obtained from the model simulation, which is then scored by calculating the least-squares error (LSE) at experimental data points k as the weighted sum over c:

$$E[\vec{p}] = \sum_c \epsilon_c \, \frac{1}{2} \sum_k (x^{dat}_{k,c} - x^{sim}_{k,c}[\vec{p}])^2 \,. \tag{9}$$

Here, ϵ_c is adjusted as to fit each combined unit comparably well to the experimental data. The same values for ϵ_c were used in the PI-SBM and the ABM and are given in Supplementary Table 1. Next, the parameter set $\vec{p}$ is randomly varied within a pre-defined neighborhood of 10% variation, leading to a new

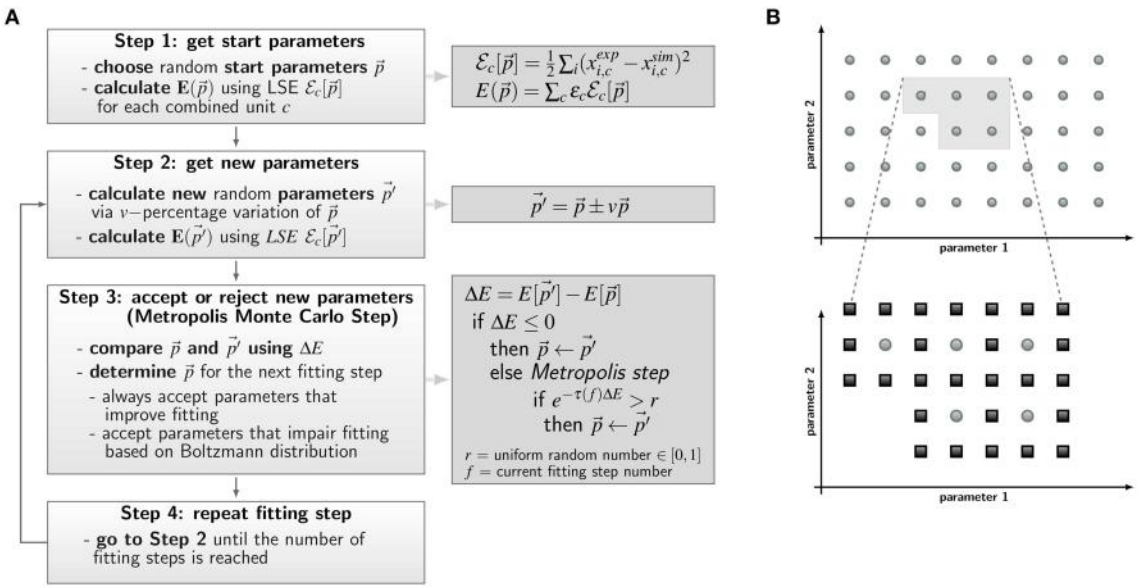


FIGURE 4 | Concepts of parameter estimation algorithms for non-spatial state-based models (SBM) and spatial agent-based models (ABM). (A) Flow chart of parameter estimation by the global optimization method Simulated Annealing based on the Metropolis Monte Carlo (SA-MMC). Light gray boxes describe Steps 1–4 of the algorithm; equations and pseudo code applied in these steps are provided in the dark gray boxes. **(B)** Schematic overview of the local optimization method *adaptive grid search algorithm* used for estimation of ABM migration parameters. Simulations are performed for sets of parameters on a regular grid in a predefined area of the search space and are evaluated by the *least squares error* (LSE). The initial grid (upper panel) contains sets of parameters with smallest LSE (light gray area) and this area is refined for a more accurate identification of the optimal parameter set (lower panel). Here, light gray dots represent parameter configurations from the first refinement level and dark gray squares represent parameter configurations of the subsequent refinement level.

set of parameter values, $\vec{p}\,'$, as indicated in **Figure 4A**, Step 2. Subsequently, the simulation of the PI-SBM is performed again for parameter values $\vec{p}\,'$ and the corresponding score $E[\vec{p}\,']$ is calculated. Whether the new simulated data will be accepted or rejected is decided by applying the MMC scheme that is depicted in **Figure 4A**, Step 3. The probability to accept worse parameter values is influenced by $\tau(f)$, representing the "inverse system temperature" in a SA process. The simulation of the annealing process involves a gradual decrease of the system temperature with progressed fitting, i.e., $\tau(f)$ is increased with the number of performed fitting steps f (see Supplementary Information 2.1).

After performing a total number of fitting steps, the fitting algorithm is repeated starting from a newly chosen random parameter set. This is done for a certain number of runs and the set of parameters with the minimal fitting error ($\vec{p}_{min}$) is saved from each fitting process. The mean values of the parameter values and their standard deviations are computed over all runs to determine the robustness of the estimated parameters.

We repeatedly perform the parameter estimation procedure for different system sizes in terms of the total number of individual cells. In doing so, the system size is stepwise increased by factors of ten, which is associated with increasing computing time for the model simulation but is partly compensated by a decrease in the number of fitting steps to avoid computational overload (see Supplementary Table 2). We start the estimation algorithm with a low number of individuals and a large number of fitting steps. The resulting parameter values are subsequently used as start parameter values for the system with next-higher number of individuals, i.e., for a 10-fold larger system. This procedure is repeated until a system size is reached where the number of individuals correspond to the measured numbers of PMN (about 5×10^6) and monocytes (about 5×10^5).

2.3.2. Adaptive Regular Grid Search

As described in Section 2.2, probabilities for state transitions in the ABM of the whole-blood infection assay can be derived from the interaction rates of the PI-SBM. This reduces the space of parameters that has to be searched in the process of parameter estimation, leaving only two migration parameters—i.e., the diffusion coefficients D_M and D_G, respectively, for monocytes and PMN—to be calibrated. However, even for a reduced parameter search space, there still is need for a calibration strategy that keeps the number of ABM simulations within limits, because simulating stochastic processes requires sufficient numbers of repetitions in order to obtain numerical results that are statistically sound.

We apply the *adaptive regular grid search algorithm* (Powell, 1998) to search iteratively for a local optimum in the parameter space (see **Figure 4B**). Motivated by biological constraints this is done for a pre-defined region of the parameter space. This region is represented on a regular grid and for each grid point the ABM is simulated with the corresponding set of parameter

values. Afterwards, simulations are evaluated with the least-squares error (LSE), scoring deviations between the simulation results and the experimental data for all combined units $c = \{C_K, C_A, C_E, C_M, C_G\}$ (see Section 2.1 and Equation 9). The values for the LSE are used to determine the adaptive refinement of the grid before the next iteration step, where intermediate grid points are calculated by bisection of the grid constant for the sets of parameters with lowest LSE. This imposes a grid refinement that ensures a more detailed scanning of the parameter space in relevant regions and defines the refinement level. The initial grid constant and the number of refinement steps determine how fine-grained the parameter space is represented by grid points and their values have to be chosen depending on the LSE landscape.

We further decrease computational costs associated with parameter estimation in the ABM by system scaling. Thus, similar to the procedure applied for the state-based model PI-SBM, we first scan the parameter space with a small system of 1/5 μl blood and subsequently re-scan the relevant parameter region with the system of 1 μl blood as defined in Section 2.2.

3. Results

3.1. Quantification of the Immune Response by the State-based Model

We quantified innate immune mechanisms in human whole-blood assays of infection with the pathogenic fungus *C. albicans* using a SBM. To this end, we modified a previously introduced SBM, referred to as P-SBM. This model was derived with the focus on state transitions of the pathogen (P) that may be induced by immune cells. However, immune cells in the P-SBM were only effectively modeled and not explicitly account for as individual cells (Hünniger et al., 2014). In the present work, we modified the P-SBM to model the interaction with individual immune cells—monocytes and granulocytes (PMN)—in detail. Since the focus of this model is on state transitions of both pathogen (P) and immune cells (I), we term this model PI-SBM. For reasons of comparison with the P-SBM, we used the same experimental data as in Hünniger et al. (2014) to quantify innate immune mechanisms by estimating the transition rates that yield the best fit to the data. Specific combinations of *C. albicans* states, referred to as *combined units*, were introduced that are directly related to different *C. albicans* populations measured over 4 h post-infection in experiment. As explained in detail in the Materials and Methods Section, the combined units include all extracellular *C. albicans* cells (C_E), *C. albicans* cells that are phagocytosed, respectively, by monocytes (C_M) or by granulocytes (C_G). Furthermore, all killed and alive *C. albicans* cells are given by the combined units C_K and C_A, respectively. The manually adjusted scores ϵ_c of combined units c are given in Supplementary Table 1. We simulate a blood sample of 1 ml containing 5×10^5 monocytes, 5×10^6 PMN and 1×10^6 *C. albicans* cells that are initially extracellular and alive.

To estimate the values of transition rates in the PI-SBM that yield the best fit to experimental data, i.e., the fit with the smallest least squares error (LSE), we applied the method of SA-MMC scheme (for details see Section 2.3.1). In **Figure 5**, the resulting transition rates of the PI-SBM are compared with those previously obtained within the P-SBM (for a quantitative comparison see also Supplementary Tables 3, 4). The direct comparison between the P-SBM and PI-SBM revealed that most transition rates are quantitatively similar in the two models.

The largest deviations in the values of transition rates between the two models were observed for the phagocytosis rate of monocytes (ϕ_M) and the killing rate of monocytes (κ_M). This was further investigated by performing the parameter estimation for the PI-SBM again, where only ϕ_M and κ_M were randomly varied while all other rates were kept fixed. We performed 50 runs and obtained very different standard deviations for these transition rates: while the standard deviation of ϕ_M was only 4%, this was 16% in the case of κ_M. We conclude that the PI-SBM is generally robust in all transition rates, except for κ_M that is also not directly determined by the data, because alive and killed *C. albicans* cells in phagocytes were not distinguished in these experiments. Similar observations were made for the P-SBM, where it was shown that variations in κ_M did not lead to significant differences in the fitting error (Hünniger et al., 2014).

To determine the impact of variations in the transition rates on the kinetics of the combined units in the PI-SBM, we performed 50 simulations with transition rates that were randomly sampled within their respective standard deviations. The kinetics of individual sub-populations are plotted in Supplementary Figure 2 while the results for the combined units are given in **Figure 6**. It can be seen that the simulated combined units agree well with the corresponding experimental data. In

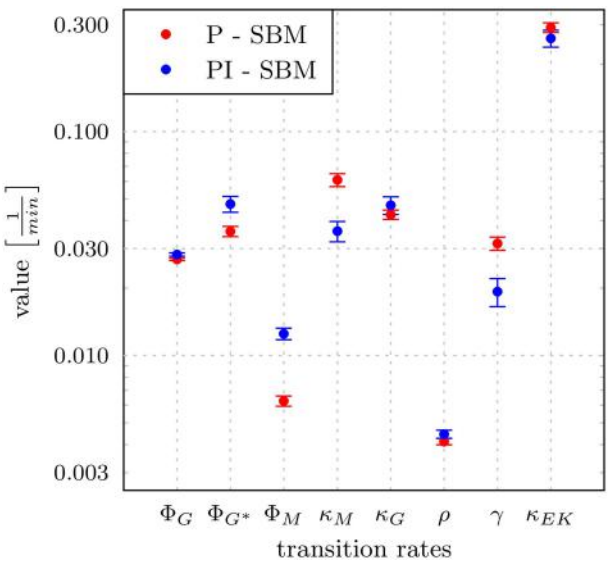


FIGURE 5 | Transition rates obtained from the model calibration to experimental data of the whole-blood infection assay. The results for the modified state-based model PI-SBM are compared to the P-SBM (Hünniger et al., 2014). The values are compared for the rate of phagocytosis by monocytes (ϕ_M), and by PMN on initial and subsequent events (ϕ_G,ϕ_{G^*}), rate of killing by monocytes (κ_M) and PMN (κ_G), rate of acquiring resistance against phagocytsis and/or killing (ρ) as well as the values of parameters for extracellular killing (γ, $\bar{\kappa}_{EK}$). Error bars correspond to standard deviations.

particular, the resulting kinetics of the PI-SBM revealed that 4 h post-infection 82% *C. albicans* cells were phagocytosed by PMN, whereas only 4% *C. albicans* cells were phagocytosed by monocytes. Furthermore, PMN play a major role in the immune response, because these phagocytes are associated with 97% of all killed *C. albicans* cells (see Supplementary Figure 2A). This is achieved either directly, via phagocytosis and intracellular killing (66.5%) of the pathogen, or indirectly by the release of antimicrobial peptides on a pathogen's first event of phagocytosis (30.5%) (see Supplementary Figure 2H). Four hours post-infection, most *C. albicans* cells were killed (89%) while a minority of 11% cells were extracellular and became resistant against killing and phagocytosis. These results are in quantitative agreement with those obtained previously for the P-SBM (Hünniger et al., 2014).

3.2. Predictions on Monocytopenia and Neutropenia from PI-SBM

The state-based model PI-SBM opens the possibility to study the dependence of the immune response against *C. albicans* on the number of PMN and monocytes in blood. Simulating the virtual infection scenario with the previously estimated parameters (see Supplementary Table 3), we considered various cases of immune cell deficiencies. The model predictions at 4 h post-infection and for gradual decreases in the immune cell numbers are presented in **Figure 7** for the cases of monocytopenia and neutropenia separately.

We found, as expected from the above quantification of the immune response, that monocytopenia is not a critical condition with regard to *C. albicans* infections: deficiency of monocytes and even their complete absence was fully compensated by PMN-mediated killing. In fact, patients with monocytopenia have not been reported to develop systemic candidiasis to date (Lionakis, 2014). The situation is extremely different in the case of neutropenia. In the absence of PMN, the number of killed *C. albicans* cells is predicted to decrease from about 89% under physiological conditions down to 45%, i.e., $C_K = 89\%$ for 5×10^6 PMN and $C_K = 45\%$ for $\leq 5 \times 10^3$ PMN (see **Figure 7B**). Monocytes compensated PMN deficiency by phagocytosis of *C. albicans* cells only partly, where the fraction increased from 3% under physiological conditions up to 48%. However, 42% of the *C. albicans* cells acquired resistance against killing and/or phagocytosis, resulting from the combined effect of absent PMN phagocytosis and extracellular killing that is normally mediated by PMN release of antimicrobial peptides.

For a decrease in PMN number by one order of magnitude from physiological conditions, we found that monocytes can sustain the immune response fairly well. In this case, the fraction of killed *C. albicans* cells was still 79% and the phagocytosis by monocytes and PMN reached, respectively, 20% and 55% of *C. albicans* cells. A significant deterioration of the immune response was observed for PMN concentrations below 5×10^5 cells/ml (see **Figure 7**). Interestingly, this concentration was reported to mark the transition from moderate to severe neutropenia (Munshi and Montgomery, 2000), which is a

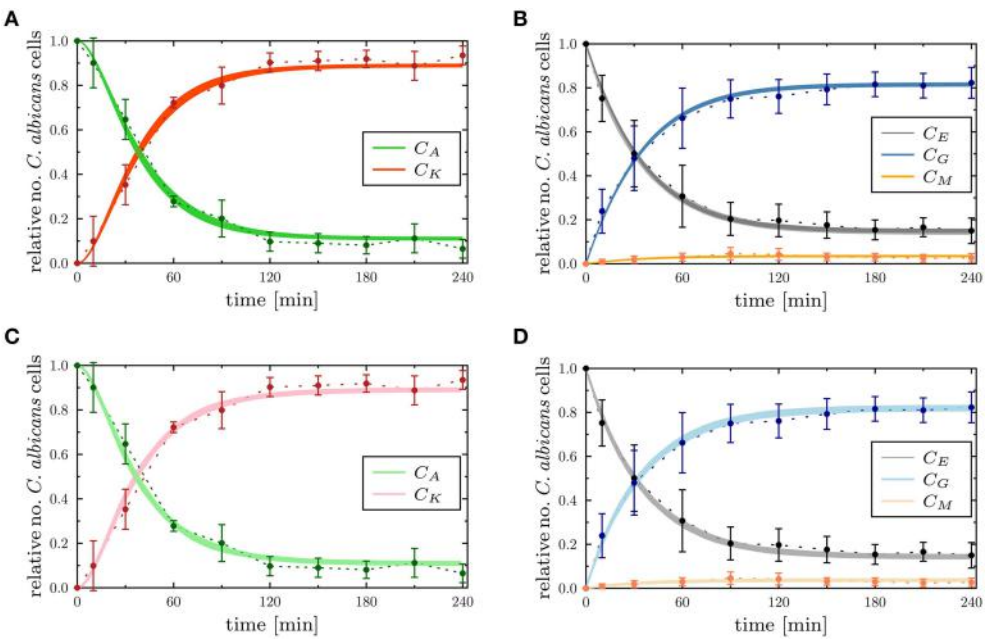


FIGURE 6 | Comparison of the time-evolution for the combined units from the experimental whole-blood infection assay (dotted lines as a guide for the eye) with the PI-SBM in (A,B), and the ABM in (C,D). In **(A,B)**, the thickness of the solid lines represents the standard deviation of the PI-SBM simulation results as obtained from 50 simulations for normally distributed transition rates as given in Supplementary Table 3. The thickness of the solid lines in **(C,D)** represents the standard deviation obtained by 30 simulations of the stochastic ABM. Time-evolution of killed (C_K) and alive (C_A) *C. albicans* cells are depicted in **(A,C)**, and the dynamics of *C. albicans* cells that are in extracellular space (C_E), phagocytosed by monocytes (C_M) and PMN (C_G) are shown in **(B,D)**.

condition that is known to be associated with high risks for candidemia in cancer patients (Lunel et al., 1999; Alangaden, 2011).

3.3. Agent-based Model Captures Immune Response in Time and Space

State-based models (SBM) do not account for any spatial aspects. For example, cells in the PI-SBM do not actually migrate during the immune response and, therefore, do not have to get into contact before a phagocytosis event can take place. In contrast, agent-based models (ABM) do capture spatial aspects in a defined environment. Applying a bottom-up modeling approach, we implemented an ABM that is—apart from its spatial aspects—the exact analog of the non-spatial PI-SBM. As depicted in **Figure 1**, all transition rates that were previously estimated for the PI-SBM were fed into the ABM (see Section 2.2 for details). The only parameters left to estimate were those related to cell migration, which in the dense cell system of the whole-blood assay resembles a random walk. In particular, while the diffusion coefficient associated with the passive movement of *C. albicans* cells could be inferred from the Stokes-Einstein equation to be $D_C \approx 1\,\mu\text{m}^2/\text{min}$, the active migration behavior of immune cells requires a rigorous parameter estimation of the diffusion coefficients D_M and D_G for monocytes and PMN, respectively.

It should be noted that, even in the case of low-dimensional parameter spaces, the estimation of parameters for ABM generally turn out to be computationally intensive. This is a consequence of the fact that ABM simulate the interactions between thousands of agents in continuous space as stochastic processes. To simultaneously facilitate an increase in model complexity and a decrease in computational expense for parameter estimation, we applied the local optimization algorithm *adaptive regular grid search*. This algorithm compares ABM simulations by evaluating the least squares error (LSE) regarding the experimental data of the whole-blood infection assay. Stochastic effects of the ABM were investigated by comparing simulation results for a fixed set of parameter values with varying number of *in silico* replicates. Using 100 *in silico* replicates as a reference for the mean value of the LSE, we generally observed for relevant parameter sets, i.e., parameter sets that yield reasonable agreement with the experimental data, that relative variations in the mean LSE were already well below 10% for 30 *in silico* replicates. Therefore, we set the number of *in silico* replicates to 30 throughout the whole parameter space.

The *adaptive regular grid search* algorithm searches the space of D_M and D_G on a pre-defined grid of diffusion coefficients, $0 < D_M, D_G < 800\,\mu\text{m}^2/\text{min}$. This range for the diffusion coefficients implies that the time step Δt_{ABM} varies between $5.2 \times 10^{-3}\,\text{min} \leq \Delta t_{\text{ABM}} \leq 4.2\,\text{min}$ (see Equation 6). As described in Section 2.3.2, we started with a relatively coarse grid of step size $100\,\mu\text{m}^2/\text{min}$ and computed at each grid point the LSE by comparing the experimental data with a small ABM system, i.e., representing $1/5\,\mu\text{l}$ of blood (see Supplementary Figure 3). These results were used to determine the regime of parameters in which the parameter estimation was continued for the large ABM system simulating $1\,\mu\text{l}$ of blood. The parameter regime was determined by the rectangle that contains all pairs of diffusion coefficients (D_G, D_M) for which the LSE values were found to be minimal from separately varying each parameter. The corner points of this rectangle were $(D_G, D_M) = (100, 0)\,\mu\text{m}^2/\text{min}$ and $(D_G, D_M) = (600, 800)\,\mu\text{m}^2/\text{min}$ (see gray region in Supplementary Figure 3). Subsequently, the grid was refined based on simulations of the large ABM by determining the path of minimal LSE values and adding grid points around this path by adaptive bisection. After simulation of the ABM for parameter sets corresponding to the added grid points, the procedure of grid refinement was repeated. This can be seen in **Figure 8**, where we plot a map of the LSE landscape together with the paths of minimal LSE values for each level of refinement. It was observed that the course of these paths covers a relatively broad range of diffusion coefficients for monocytes, D_M, whereas this is a fairly narrow range of D_G-values for PMN.

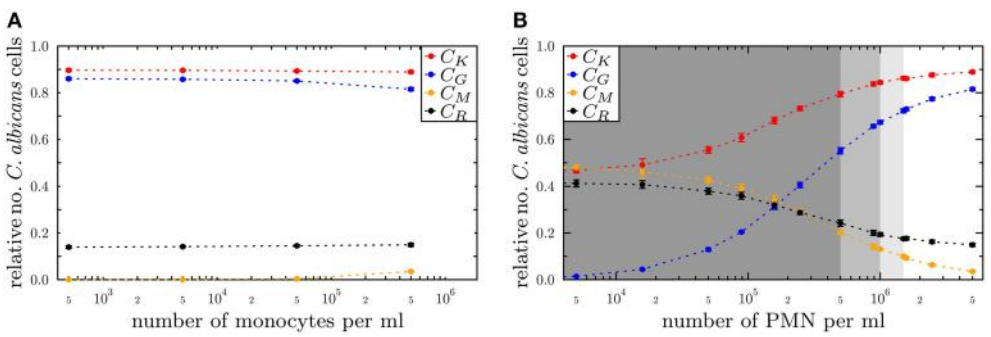


FIGURE 7 | Simulation results of the PI-SBM with different immune cell numbers at 4 h post-infection for the conditions (A) monocytopenia and (B) neutropenia. The relative numbers of *C. albicans* cells of killed cells (C_K), phagocytosed cells in monocytes (C_M) and in PMN (C_G) as well as cells that became resistant (C_R) against killing and/or phagocytosis are depicted for different numbers of monocytes and PMN. The number of **(A)** monocytes and **(B)** PMN in the simulations are reduced separately, starting from physiological concentrations of $5{\times}10^5$ /*ml* monocytes and $5{\times}10^6$ /ml PMN down to vanishing concentrations. In **(B)**, the light gray region represents the range of light neutropenia ($< 1.5 \times 10^6$ PMN per ml), medium gray region represents the range of moderate neutropenia ($< 1 \times 10^6$ PMN per ml) and dark gray region represents the range of severe neutropenia ($< 5 \times 10^5$ PMN per ml).

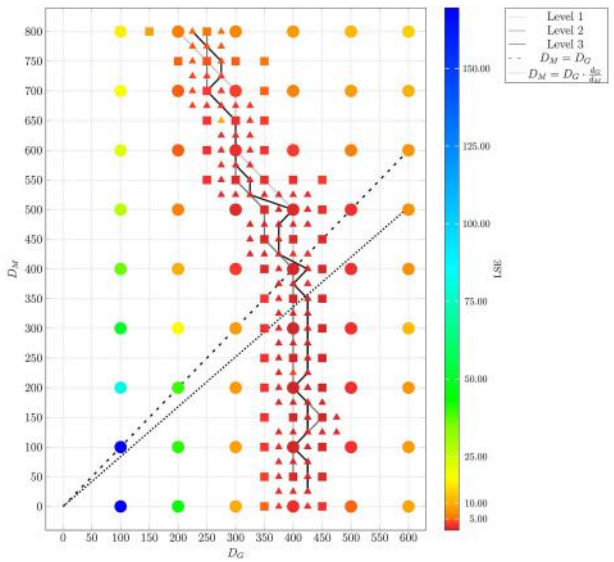


FIGURE 8 | Result of ABM parameter estimation by *adaptive regular grid search*. The diffusion coefficients for monocytes, D_M, and PMN, D_G, were scanned within the regime that was determined by parameter scanning for the small ABM (1/5 μl blood). At each grid point, 30 ABM simulations were performed for the large system (1 μl blood) and the mean least squares error (LSE) is depicted. By determining the path of minimal LSE values and adding grid points around this path by adaptive bisection, three refinement levels are considered. Dots represent grid points of the first refinement level, squares represent grid points of the second refinement level, and triangles represent grid points of the third refinement level. The paths of minimal LSE values for the first, second and third refinement level are traced by light gray, medium gray, and dark gray lines, respectively.

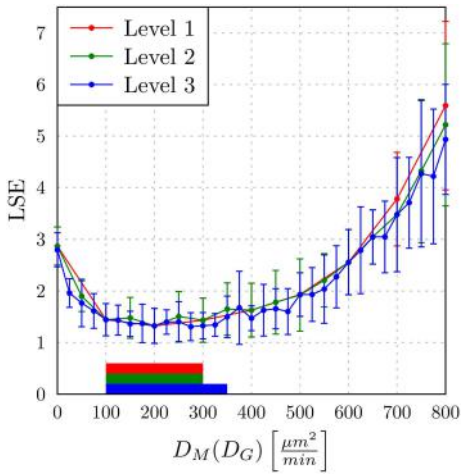


FIGURE 9 | The least squares error (LSE) of the paths of grid points along the diffusion coefficients for monocytes (D_M) as a function of the minimal diffusion coefficient for PMN (D_G): $D_M(D_G)$. Mean values and standard deviations were obtained from averaging over 30 ABM simulations. The paths of the first, second, and third refinement level are shown, respectively, as red, green, and blue lines (guide for the eyes). The horizontal bars indicate regions of diffusion coefficients with values comparable to the absolute LSE minimum of each refinement level. All values outside these regions are significantly different from the absolute LSE minimum (Wilcoxon rank sum test with $p < 0.05$).

In **Figure 9**, we present the LSE values as a function of $D_M(D_G)$ along the paths of minimal LSE values for the three levels of refinement. From the third level of refinement we inferred the point of absolute LSE minimum to be located at $(D_G^{min}, D_M^{min}) = (425, 275)\,\mu\text{m}^2/\text{min}$. However, since the landscape of $D_M(D_G)$ resembled an extended valley across neighboring data points, we performed a statistical analysis by applying the Wilcoxon rank sum test between the absolute LSE minimum and its neighboring points to check for significant differences between them. Imposing a p-value of $p < 0.05$ for significant difference, we obtained points with similar values of the LSE ranging in the interval $D_M = 100\,\mu\text{m}^2/\text{min}$ to $D_M = 350\,\mu\text{m}^2/\text{min}$ for monocytes and $D_G = 400\,\mu\text{m}^2/\text{min}$ to $D_G = 425\,\mu\text{m}^2/\text{min}$ for PMN (see **Figure 8**). These findings imply that the immune response in the whole-blood infection assay was much more sensitive to variations in the diffusion coefficients of PMN than of monocytes, emphasizing the dominant role of PMN over monocytes from the viewpoint of cell migration.

Our results are consistent with the absolute LSE minima of refinement level one and two, which were both at $(D_G^{min}, D_M^{min}) = (400, 200)\,\mu\text{m}^2/\text{min}$ and that also belong to this interval (see **Figure 9**). Interestingly, while we expected that monocytes are less migratory active than PMN, i.e., restricting the relevant parameter regime in **Figure 8** to the region below the dashed line, we also found that the interval around the absolute LSE minimum contains the parameter set $(D_G, D_M) = (425, 350)\,\mu\text{m}^2/\text{min}$. The ratio of these diffusion coefficients, $D_M/D_G \approx 0.82$, resembles the value expected from the Stokes-Einstein equation (Einstein, 1905) implying $D_M/D_G = d_G/d_M$ (dotted line in **Figure 8**). Taken together, we consider the diffusion coefficients $(D_G^{min}, D_M^{min}) = (425, 275)\,\mu\text{m}^2/\text{min}$ to represent the immune cell dynamics reasonably well and use these values in our further analyses below.

Next, we compared the ABM simulation results for the absolute LSE minimum with those of the PI-SBM. These are plotted together with the experimental data of the whole-blood infection assay in **Figure 6** and in Supplementary Figure 4 for the time evolution of *C. albicans* sub-populations. Thus, we found that both modeling approaches, the non-spatial SBM and the spatial ABM, yielded excellent agreement with the experimental data. Furthermore, we found that our simulation results obtained from the stochastic ABM were robust, which can be seen from the line thicknesses in **Figures 6C,D** representing the standard deviations obtained from 30 ABM simulations.

3.4. Predictions on Hyper- and Hypo-inflammation from ABM

To investigate the impact of hyper- and hypo-inflammation associated with the dynamics of immune cells, we varied the diffusion coefficients of monocytes and PMN separately around the absolute LSE minimum $(D_G^{min}, D_M^{min}) = (425, 275)\,\mu\text{m}^2/\text{min}$. Keeping the diffusion coefficient D_G fixed and varying the D_M for monocytes between $100\,\mu\text{m}^2/\text{min}$ and $600\,\mu\text{m}^2/\text{min}$, we

observed at 4 h post-infection no substantial changes in the populations of killed, resistant and phagocytosed *C. albicans* cells (see **Figure 10A**). At extreme values $D_M > D_G$, a slight increase (decrease) in the number of killed (resistant) *C. albicans* cells was observed accompanied by a slight increase in the phagocytosis by both monocytes and PMN. This may be attributed to a stronger mixing of the cell system for high diffusion coefficients D_M. In general, however, the immune response does not appear to be sensitive to this parameter, which is in agreement with the finding for monocytopenia that did not have a substantial impact on infection clearance (see **Figure 7A**).

In the opposite case, where D_M was fixed and D_G was varied between 100 $\mu m^2/min$ and 600 μm^2/min, it was again observed that for increased values $D_G > 425\,\mu$m^2/min the impact on the immune response against *C. albicans* is only weak. In contrast, for decreased values $D_G < 400\,\mu$m^2/min the immune response was strongly affected by the reduced migratory activity of PMN. This could be observed by a substantial increase (decrease) in the number of resistant (killed) *C. albicans* cells (see **Figure 10B**). In particular, for $D_G = 100\,\mu$m^2/min the phagocytosis of *C. albicans* cells by PMN was reduced by more than 20% and the relative number of resistant *C. albicans* cells reached the value of 28%. Comparing this scenario with the condition of PMN deficiency (see **Figure 7B**), we found that this PMN paralysis resembles moderate to severe neutropenia associated with a relative number of about 20% and 30% of resistant *C. albicans* cells, respectively.

4. Discussion

In this study, we applied a bottom-up modeling approach to simulate an experimental infection assay for *C. albicans* in human blood. As illustrated in **Figure 1**, this approach combines different mathematical models with increasing complexity that build on one another. We started from a previously developed state based model (SBM), here referred to as P-SBM (Hünniger et al., 2014), that neglects all spatial aspects and describes the time-evolution of pathogens in different states—e.g., alive, phagocytosed and killed—during the early response of the innate immune system. To include the immune response mediated by monocytes and granulocytes (PMN), in this work we modified the P-SBM into a SBM that does as well-explicitly account for the immune cell states and is therefore referred to as PI-SBM. The rates of state transitions in the PI-SBM were estimated by comparison with experimental data (Hünniger et al., 2014) using the global optimization method *simulated annealing* based on the Metropolis Monte Carlo scheme (SA-MMC).

The resulting model kinetics of both SBM were found to be in quantitative agreement with experimental data and confirmed that PMN play the major role in the immune defense against *C. albicans* in human blood. This is indicative for the general validity of both models, despite the structural difference of the simulation algorithms regarding the level of detail at which immune cells are modeled. Furthermore, the PI-SBM allows making predictions on infection scenarios in patients with immune cell deficiencies, i.e., neutropenia and monocytopenia. Performing *in silico* experiments with varying numbers of either monocytes or PMN, revealed that loss of monocytes was mainly compensated by PMN. In contrast, decreasing PMN number lead to a strongly reduced immune response against *C. albicans* for PMN numbers below 5×10^5 /ml (see **Figure 7**). Our quantitative prediction is substantiated by published data that account this PMN concentration as severe neutropenia (Munshi and Montgomery, 2000). It is also reported that neutropenia impairs the outcome of candidemia and is a risk factor, in particular, for cancer patients developing candidemia (Guiot et al., 1994; Bow et al., 1995; Lunel et al., 1999). From the quantitative agreement between predictions of the PI-SBM and reported findings for *C. albicans* infection, we attribute a high predictive potential to this virtual infection model that may be exploited in future studies, e.g., focusing on conditions of immune dysregulation and/or comparing the impact of different pathogens. The possibility to quantify functional alteration of immune cells rather than pure numerical aberrations is of particular interest in this regard.

In order to account for spatial aspects of the immune response, we extended the SBM to an agent-based model (ABM), where cells are simulated as agents that can migrate in continuous

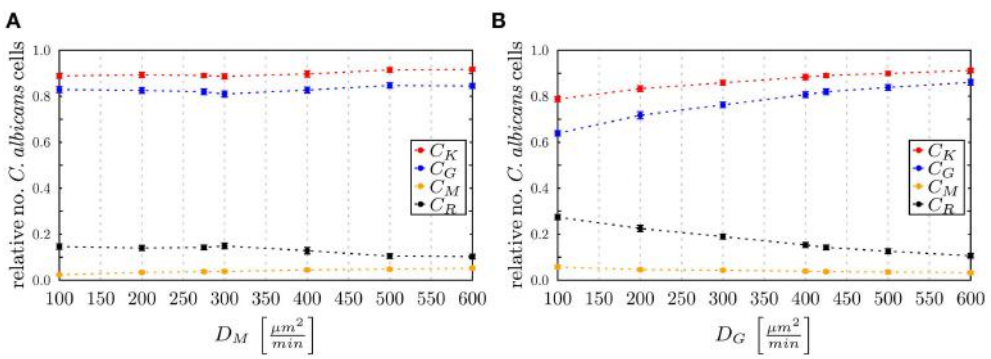


FIGURE 10 | Simulation results of the ABM at 4 h post-infection for varied diffusion coefficients around the absolute least squares error (LSE) minimum with (D_G^{min}, D_M^{min}) = (425, 275) μm^2/min for (A) monocytes keeping D_G fixed and (B) PMN keeping D_M fixed. The relative numbers of *C. ablicans* cells of killed cells (C_K), phagocytosed cells in monocytes (C_M) and in PMN (C_G) as well as cells that became resistant (C_R) against killing and/or phagocytosis are depicted.

three-dimensional space and can interact with each other on encounter in space. Applying the bottom-up modeling approach, we made use of the rates that were determined by fitting the PI-SBM to the experimental data and estimated the diffusion coefficients of immune cells in blood (see **Figure 1**). Due to high computational costs of ABM simulations, applying the global optimization method SA-MMC was no realistic option and we chose the computationally affordable local optimization method *adaptive regular grid search*. This method searches for the optimum within a pre-defined parameter space, which in the present case was a two-dimensional space spanned by the diffusion coefficients for monocytes and PMN. In contrast, applying SA-MMC was beneficial in the case of PI-SBM for three reasons: (i) the parameter space was eight-dimensional, (ii) limitations of the parameter space would have been difficult to motivate biologically, and (iii) computational costs for repeated simulations were still acceptable due to the neglect of spatial aspects.

As live cell imaging in whole-blood assays cannot yet be performed today, computer simulations are the only tool to predict diffusion coefficients of immune cells. Parameter estimation of the ABM predicted intervals for the diffusion coefficients that yielded quantitatively comparable results. For monocytes this interval, $D_M = 100\,\mu\text{m}^2/\text{min}$ to $D_M = 350\,\mu\text{m}^2/\text{min}$, was substantially broader than for PMN with $D_G = 400\,\mu\text{m}^2/\text{min}$ to $D_G = 425\,\mu\text{m}^2/\text{min}$, indicating the importance of fine-tuned PMN motility.

Furthermore, by varying the diffusion coefficients of the immune cells, we demonstrated the impact of hyper- and hypo-inflammation in immune dysregulation. In general, reducing (increasing) immune cell motilities around optimal values reduced (increased) the number of interaction events between cells and by that the phagocytosis of *C. albicans* cells. In the case of PMN, reduction of cell motility and phagocytosis events was additionally associated with a decrease in the release of antimicrobial peptides contributing to the decrease in killing of *C. albicans* cells. This in turn lead to an increase in the number of resistant *C. albicans* cells reaching levels that were well-beyond those observed for paralytic monocytes (see **Figure 10**). Comparing the hypo-inflammatory condition with PMN deficiency, we found that diffusion coefficients around $D_G = 100\,\mu\text{m}^2/\text{min}$ resembled the outcome of moderate to severe neutropenia.

The bottom-up modeling approach presented here may be extended in various ways. For example, the implementation of a hybrid ABM could be envisaged where molecular interactions, e.g., as mediated by antimicrobial peptides, are not simulated in a rule-based fashion but in an explicit way by a molecular diffusion solver. Other directions of future research include (i) focusing on conditions of immune dysregulation, (ii) comparing the impact of different pathogens, and (iii) including other types of innate immune cells. Furthermore, it is conceivable to combine modeling approaches with microscopy experiments of infection scenarios *in vitro* in an image-based systems biology approach (Mech et al., 2014; Figge and Murphy, 2015; Medyukhina et al., 2015). First steps into this direction have recently been made, e.g., by establishing algorithms for the automated image analysis of phagocytosis assays (Mech et al., 2011; Kraibooj et al., 2014) and for the automated tracking and classification of PMN from time-lapse microscopy (Mokhtari et al., 2013; Brandes et al., 2015) that was applied in the context of comparing *C. albicans* and *C. glabrata* infection (Duggan et al., 2015a). In the future, we expect that a systems medicine approach exploiting the predictive power of virtual infection models will play an important role in the context of infectious disease diagnosis.

Author Contributions

TL, ST, MTF: Conception and design of the investigation and work. MTF: Contribution of materials and computational resources. TL, ST, JP, MTF: Data processing, implementation and application of the computational algorithm. TL, ST, JP, KH, OK, MTF: Evaluation and analysis of the results. TL, ST, JP, KH, OK, MTF: Drafting the manuscript and revising it critically for important intellectual content and final approval of the version to be published. TL, ST, JP, KH, OK, MTF: Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding

This work was financially supported by the Deutsche Forschungsgemeinschaft (DFG) through the excellence graduate school Jena School for Microbial Communication (JSMC) and the CRC/TR124 FungiNet (project B4 to MTF and project C3 to OK).

Acknowledgment

We thank C. M. Svensson for valuable discussions on the statistical analysis of the agent-based model simulations.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00608/abstract

References

Alangaden, G. J. (2011). Nosocomial fungal infections: epidemiology, infection control, and prevention. *Infect. Dis. Clin. North Am.* 25, 201–225. doi: 10.1016/j.idc.2010.11.003

Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J. A., and Blom, J. G. (2009). Systems biology: parameter estimation for biochemical models. *FEBS J.* 276, 886–902. doi: 10.1111/j.1742-4658.2008.06844.x

Bittig, A. T., and Uhrmacher, A. M. (2010). "Spatial modeling in cell biology at multiple levels," in *Simulation Conference (WSC),*

Proceedings of the 2010 Winter, Number 2005 (Baltimore, MD: IEEE), 608–619.

Bow, E. J., Loewen, R., Cheang, M. S., and Schacter, B. (1995). Invasive fungal disease in adults undergoing remission-induction therapy for acute myeloid leukemia: the pathogenetic role of the antileukemic regimen. *Clin. Infect. Dis.* 21, 361–369.

Brandes, S., Mokhtari, Z., Essig, F., Hünniger, K., Kurzai, O., and Figge, M. T. (2015). Automated segmentation and tracking of non-rigid objects in time-lapse microscopy videos of polymorphonuclear neutrophils. *Med. Image Anal.* 20, 34–51. doi: 10.1016/j.media.2014.10.002

Cunha, C., Kurzai, O., Löffler, J., Aversa, F., Romani, L., and Carvalho, A. (2014). Neutrophil responses to aspergillosis: new roles for old players. *Mycopathologia* 178, 387–393. doi: 10.1007/s11046-014-9796-7

Dueck, G. (1993). New optimization heuristics. *J. Comput. Phys.* 104, 86–92.

Duggan, S., Essig, F., Hünniger, K., Mokhtari, Z., Bauer, L., Lehnert, T., et al. (2015a). Neutrophil activation by *Candida glabrata* but not *Candida albicans* promotes fungal uptake by monocytes. *Cell. Microbiol.* doi: 10.1111/cmi.12443. [Epub ahead of print].

Duggan, S., Leonhardt, I., Hünniger, K., and Kurzai, O. (2015b). Host response to *Candida albicans* bloodstream infection and sepsis. *Virulence*. doi: 10.4161/21505594.2014.988096. [Epub ahead of print].

Einstein, A. (1905). Über die von der molekularkinetischen theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.* 322, 549–560.

Fahraus, R., and Lindqvist, T. (1931). The viscosity of the blood in narrow capillary tubes. *Am. J. Physiol.* 96, 562–568.

Figge, M. T., and Murphy, R. (eds.). (2015). Image-based systems biology. *Cytometry A*. 87, 459–461. doi: 10.1002/cyto.a.22638

Figge, M. T. (2005). Stochastic discrete event simulation of germinal center reactions. *Phys. Rev. E* 71:051907. doi: 10.1103/PhysRevE.71.051907

Gonzalez, O. R., Kueper, C., Jung, K., Naval, P. C., and Mendoza, E. (2007). Parameter estimation using simulated annealing for S-system models of biochemical networks. *Bioinformatics* 23, 480–486. doi: 10.1093/bioinformatics/btl522

Guiot, H., Fibbe, W., and Van't Wout, J. (1994). Risk factors for fungal infection in patients with malignant hematologic disorders: implications for empirical therapy and prophylaxis. *Clin. Infect. Dis.* 18, 525–532.

Hünniger, K., Lehnert, T., Bieber, K., Martin, R., Figge, M. T., and Kurzai, O. (2014). A virtual infection model quantifies innate effector mechanisms and *Candida albicans* immune escape in human blood. *PLoS Comput. Biol.* 10:e1003479. doi: 10.1371/journal.pcbi.1003479

Hünniger, K., Bieber, K., Martin, R., Lehnert, T., Figge, M. T., Löffler, J., et al. (2015). A second stimulus required for enhanced antifungal activity of human neutrophils in blood is provided by Anaphylatoxin C5a. *J. Immunol.* 194, 1199–1210. doi: 10.4049/jimmunol.1401845

Hernandez-Vargas, E. A., Wilk, E., Canini, L., Toapanta, F. R., Binder, S. C., Uvarovskii, A., et al. (2014). The effects of aging on influenza virus infection dynamics. *J. Virol.* 88, 4123–4131. doi: 10.1128/JVI.03644-13

Horn, F., Heinekamp, T., Kniemeyer, O., Pollmächer, J., Valiante, V., and Brakhage, A. A. (2012). Systems biology of fungal infection. *Front. Microbiol.* 3:108. doi: 10.3389/fmicb.2012.00108

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671–680.

Kraibooj, K., Park, H.-R., Dahse, H.-M., Skerka, C., Voigt, K., and Figge, M. T. (2014). Virulent strain of lichtheimia corymbifera shows increased phagocytosis by macrophages as revealed by automated microscopy image analysis. *Mycoses* 57, 56–66. doi: 10.1111/myc.12237

Lionakis, M. S. (2014). New insights into innate immune control of systemic candidiasis. *Med. Mycol.* 52, 555–564. doi: 10.1093/mmy/myu029

Lunel, F. M. V., Meis, J. F., and Voss, A. (1999). Nosocomial fungal infections: candidemia. *Diagn. Microbiol. Infect. Dis.* 34, 213–220.

Luo, S., Skerka, C., Kurzai, O., and Zipfel, P. F. (2013). Complement and innate immune evasion strategies of the human pathogenic fungus *Candida albicans*. *Mol. Immunol.* 56, 161–169. doi: 10.1016/j.molimm.2013.05.218

Mak, T. W., and Saunders, M. E. (2011). *Primer to the Immune Response: Academic Cell Update Edition*. Burlington; San Diego; London: Academic Press.

Margulies, L., and Schwartz, K. V. (1998). *Five Kingdoms - An Illustrated Guide to the Phyla of Life on Earth, 3rd Edn.* New York, NY: W. H. Freeman and Company.

McClatchey, K. D. (2003). Clinical laboratory medicine. *Clin. Chem.* 49, 344–345. doi: 10.1373/49.2.344

Mech, F., Thywißen, A., Guthke, R., Brakhage, A. A., and Figge, M. T. (2011). Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS ONE* 6:e19591. doi: 10.1371/journal.pone.0019591

Mech, F., Wilson, D., Lehnert, T., Hube, B., and Thilo Figge, M. (2014). Epithelial invasion outcompetes hypha development during *Candida albicans* infection as revealed by an image-based systems biology approach. *Cytometry A* 85, 126–139. doi: 10.1002/cyto.a.22418

Medyukhina, A., Timme, S., Mokhtari, Z., and Figge, M. T. (2015). Image-based systems biology of infection. *Cytometry A* 87, 462–470. doi: 10.1002/cyto.a.22638

Mendling, W. (2006). *Vaginose, Vaginitis und Zervizitis*. Heidelberg: Springer Science & Business Media.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087.

Mokhtari, Z., Mech, F., Zitzmann, C., Hasenberg, M., Gunzer, M., and Figge, M. T. (2013). Automated characterization and parameter-free classification of cell tracks based on local migration behavior. *PLoS ONE* 8:e80808. doi: 10.1371/journal.pone.0080808

Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* 13, 2467–2474. doi: 10.1101/gr.1262503

Munshi, H. G., and Montgomery, R. B. (2000). Evidence-based case review: severe neutropenia: a diagnostic approach. *West. J. Med.* 172, 248–252. doi: 10.1136/ewjm.172.4.248

Pollmächer, J., and Figge, M. T. (2014). Agent-based model of human alveoli predicts chemotactic signaling by epithelial cells during early *Aspergillus fumigatus* infection. *PLoS ONE* 9:e111630. doi: 10.1371/journal.pone.0111630

Powell, M. (1998). Direct search algorithms for optimization calculations. *Acta Numerica* 7, 287–336.

Rapaport, D. C. (1996). *The Art of Molecular Dynamics Simulation*. New York, NY: Cambridge University Press.

Rosenson, R., McCormick, A., and Uretz, E. (1996). Distribution of blood viscosity values and biochemical correlates in healthy adults. *Clin. Chem.* 42, 1189–1195.

Skvoretz, J. (2002). Complexity theory and models for social networks. *Complexity* 8, 47–55. doi: 10.1002/cplx.10062

Storn, R., and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Opt.* 11, 341–359.

Tokarski, C., Hummert, S., Mech, F., Figge, M. T., Germerodt, S., Schroeter, A., et al. (2012). Agent-based modeling approach of immune defense against spores of opportunistic human pathogenic fungi. *Front. Microbiol.* 3:129. doi: 10.3389/fmicb.2012.00129

Von Neumann, J. (1951). The general and logical theory of automata. *Cereb. Mech. Behav.* 1, 41.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 28 May 2015
doi: 10.3389/fmicb.2015.00503

Deciphering chemokine properties by a hybrid agent-based model of *Aspergillus fumigatus* infection in human alveoli

Johannes Pollmächer[1,2] *and Marc Thilo Figge*[1,2]*

[1] *Applied Systems Biology, Leibniz-Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany,* [2] *Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, Jena, Germany*

The ubiquitous airborne fungal pathogen *Aspergillus fumigatus* is inhaled by humans every day. In the lung, it is able to quickly adapt to the humid environment and, if not removed within a time frame of 4–8 h, the pathogen may cause damage by germination and invasive growth. Applying a to-scale agent-based model of human alveoli to simulate early *A. fumigatus* infection under physiological conditions, we recently demonstrated that alveolar macrophages require chemotactic cues to accomplish the task of pathogen detection within the aforementioned time frame. The objective of this study is to specify our general prediction on the as yet unidentified chemokine by a quantitative analysis of its expected properties, such as the diffusion coefficient and the rates of secretion and degradation. To this end, the rule-based implementation of chemokine diffusion in the initial agent-based model is revised by numerically solving the spatio-temporal reaction-diffusion equation in the complex structure of the alveolus. In this hybrid agent-based model, alveolar macrophages are represented as migrating agents that are coupled to the interactive layer of diffusing molecule concentrations by the kinetics of chemokine receptor binding, internalization and re-expression. Performing simulations for more than a million virtual infection scenarios, we find that the ratio of secretion rate to the diffusion coefficient is the main indicator for the success of pathogen detection. Moreover, a subdivision of the parameter space into regimes of successful and unsuccessful parameter combination by this ratio is specific for values of the migration speed and the directional persistence time of alveolar macrophages, but depends only weakly on chemokine degradation rates.

Keywords: ***Aspergillus fumigatus*, fungal infection, agent-based modeling, reaction-diffusion equation, chemotaxis, human alveolus, alveolar macrophage, alveolar epithelial cell**

OPEN ACCESS

Edited by:
Tunahan Cakir,
Gebze Technical University, Turkey

Reviewed by:
Joachim Selbig,
University of Potsdam, Germany
Mark Read,
The University of Sydney, Australia

***Correspondence:**
Marc Thilo Figge,
Applied Systems Biology,
Leibniz-Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Adolf-Reichwein-Str. 23, Jena 07749, Germany
thilo.figge@hki-jena.de

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *19 February 2015*
Accepted: *06 May 2015*
Published: *28 May 2015*

Citation:
Pollmächer J and Figge MT (2015) Deciphering chemokine properties by a hybrid agent-based model of Aspergillus fumigatus infection in human alveoli. Front. Microbiol. 6:503. doi: 10.3389/fmicb.2015.00503

1. Introduction

Aspergillus fumigatus is the most dangerous airborne fungal pathogen in humans leading to high mortality rates (Heinekamp et al., 2014). Immunocompetent individuals are able to prevail over inhaled conidia of the fungus in an everyday challenge. In contrast, patients with an altered immune system, e.g., as a consequence of organ transplantation or an underlying disease like HIV, are at high risk to die from invasive aspergillosis (Horn et al., 2012), where the lung is the site of infection

in 70 % of the cases (Lin et al., 2001). *A. fumigatus* is able to adapt within hours to the humid and nutrient rich milieu of the lung (Hohl, 2008; Hasenberg et al., 2011), by this setting a tight time scale for phagocytes to find, detect and remove the pathogenic fungus before the onset of germination and hyphal invasion of alveolar epithelium.

Alveolar macrophages (AM) reside on the inner surface of lung alveoli and are the first professional motile phagocytes that get in contact with inhaled conidia of *A. fumigatus* (Hasenberg et al., 2013). AM are capable of clearing the lower respiratory tract from all kinds of inhaled particles and microbes in order to maintain a pathogen-free alveolar surface and to ensure optimal exchange of oxygen and carbon-dioxide (Fels and Cohn, 1986). The migration of AM takes place within the alveolar lining layer, which is a viscous fluid—referred to as surfactant—that coats the alveolar surface with an average thickness of about 200 nm (Bastacky and Lee, 1995). Apart from the stabilizing effect of the surfactant avoiding alveolus collapse, it also provides the environment for diffusive transport of molecules, such as lipids and immunoregulatory proteins SP-A and SP-D, that are continuously produced, secreted and recycled by alveolar epithelial cells (AEC) (Herzog et al., 2008).

For over one decade computational approaches have proven to successfully complement wet-lab studies in the frame of systems biology (Kitano, 2002; Horn et al., 2012). Computer modeling and simulation are nowadays important tools to verify hypotheses in advance of cost- and time intensive experimental investigations to narrow down the range of possible wet-lab experiments to the most promising ones. Furthermore, predictions may be derived from virtual models, which subsequently can be tested in experiment. The present study aims at predicting AM chemokine properties from an existing agent-based virtual infection model of human alveoli under physiological conditions (Pollmächer and Figge, 2014). Due to the peculiar physiology of the human lung, investigations *in vivo*, including live-cell imaging, are hard to realize. Thus, quantitative measures like AM motility, chemokine secretion rates of AEC or the diffusion coefficient of molecules within the surfactant are not directly accessible. AEC type II cell lines have been studied intensively in the past, but as they do account for only five percent of the alveolar surface, experimental investigations of type I AEC would be highly appreciated. However, isolation and cultivation of type I AEC with current methods are demanding tasks due to their thin and delicate morphology. The present computational modeling approach enables us studying *A. fumigatus* infection in alveoli for varying parameter sets of AM motility and of chemokine properties in order to reveal the relative importance of each of the parameters and their potential regimes in healthy individuals.

Recently, we established an agent-based model (ABM) of *A. fumigatus* infection in the human alveolus to study the early immune response under physiological conditions (Pollmächer and Figge, 2014). In this three-dimensional to-scale model, we represented the human alveolus by a three-quarter spherical structure consisting of type I and type II AEC as well as pores of Kohn. Our computations of the first-passage-time, i.e., the time it takes until the conidium is detected by an AM for the first time, clearly showed that pathogen detection by AM resembles the problem of finding the needle in the haystack within a time limit that is set by the germination time for *A. fumigatus* conidia of about 6 h. Statistical analyses based on hundreds of thousands of computer simulations revealed that for AM to successfully accomplish finding the conidium within 6 h time, chemotactic cues are required that guide AM to the AEC associated with a conidium. Chemotaxis was implemented in the ABM based on a probabilistic rule, i.e., AM were directed toward the AEC associated with the fungus with a probability that was defined by the distance-dependent strength of the chemokine gradient (Pollmächer and Figge, 2014). The gradient of the chemokine concentration in the alveolus was approximated by the analytical steady state solution of the two-dimensional diffusion equation for a point source on a planar surface. We demonstrated that this level of detail was sufficient to arrive at the conclusion that chemotactic cues are required for directing AM migration in the alveolus to the site of the pathogen. However, the specific chemokine remains as of yet unknown, including its characteristic parameters such as the secretion rate, diffusion coefficient and rate of degradation. In order to arrive at quantitative predictions of characteristic parameters that may narrow down the regime of candidate chemokines, the ABM has to be revised to describe the spatio-temporal dynamics of chemokine diffusion in the alveolus and the receptor binding on AM at a higher level of detail.

Mathematical models of chemotaxis typically set focus on one of the three key aspects that are associated with the directed migration of cells: gradient-sensing, polarization and motility. While integrative models combining all three aspects are still rare today (Iglesias and Devreotes, 2008), a chemotaxis model including the processes of gradient-sensing and motility was developed by Guo and Tay (2008). In this approach, a hybrid ABM (hABM) was used to simulate the migration behavior of leucocytes and to compare with experimental results of under-agarose assays. A hABM is a multi-scale model where cells are represented as migrating and interacting agents that are coupled to the interactive layer of diffusing molecule concentrations by the kinetics of chemokine receptor binding, internalization and re-expression (see **Figure 1**). From a technical point of view, this requires the implementation of a solver for the spatio-temporal reaction-diffusion equation of molecule concentrations in the complex alveolar structure with spherical symmetry and peculiar boundary conditions as imposed by the pores of Kohn and the alveolar entrance ring. This is achieved by generating a Delaunay triangulation of the alveolar surface for close-to-equidistant surface points. The geometric quantities of the corresponding Voronoi tesselation, i.e., the dual graph of the Delaunay triangulation, can then be used to solve the reaction-diffusion equation by a finite difference method on unstructured grids (Sukumar, 2003). We perform a numerical study of the steady state behavior of molecules for typical values of the diffusion coefficient, chemokine secretion rate and the rate of molecular degradation. Furthermore, performing statistical analyses of first-passage-time distributions we narrow down the regime of characteristic parameters required for the time-limited detection of *A. fumigatus* conidia by AM.

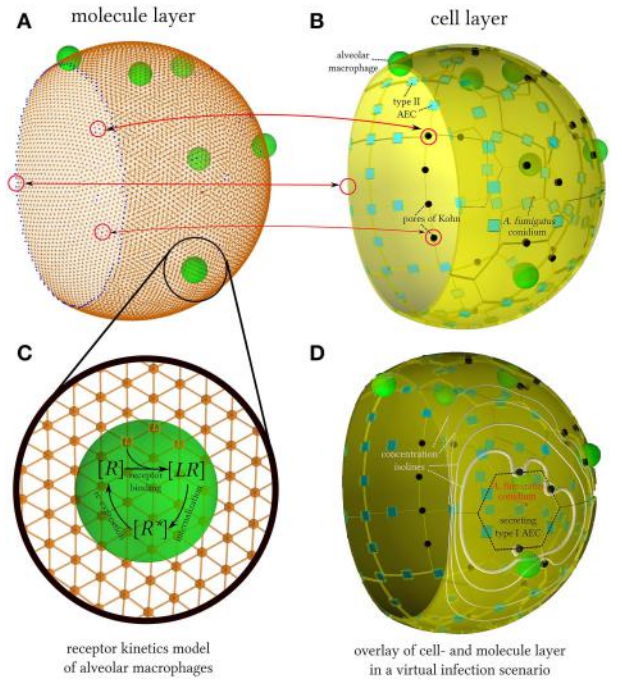


FIGURE 1 | Schematic overview and structural relations between different components of the hybrid agent-based model. (A) Close-to-equidistant discretization of the three-quarter alveolus with 10,000 grid points. Grid points with label alveolar surface point (*orange spheres*) are connected with their neighboring grid points (*orange lines*) and those with label boundary point (*blue spheres*) correspond to either pores of Kohn or the alveolar entrance ring. **(B)** To-scale reconstruction of the human three-quarter alveolus from Pollmächer and Figge (2014) including alveolar epithelial cells (AEC) of type I (*yellow*) and type II (*blue*) as well as the pores of Kohn (*black*). **(C)** Receptor kinetics model that drives the chemotaxis of alveolar macrophages (AM). Free chemokine receptors [*R*] bind to chemokine ligands [*L*] located at grid points associated with the AM. Bound receptors [*LR*] are processed into internalized receptors [*R**] and are re-expressed subsequently. **(D)** Snapshot of a virtual infection scenario where AM (*green*) aim to find a conidium of *A. fumigatus* (*red*). The information contained in the molecule layer is integrated using chemokine concentration isolines (*white*), which are plotted proportional to their respective values with different sizes.

2. Materials and Methods

2.1. Hybrid Agent-Based Model

In this study, we revised our agent-based model (ABM) of the human alveolus to explicitly account for the dynamics of molecular diffusion and reactions with cells, which were previously modeled in a simple rule-based fashion using a steady-state approximation (Pollmächer and Figge, 2014). We refer to the revised model as hybrid agent-based model (hABM), because single cells are represented as individual agents that migrate and interact in continuous space, whereas chemokine concentrations are represented as spatio-temporal distributions on a discrete grid. In this multi-scale approach, interactions between cellular agents and the layer of diffusing molecular concentrations are realized via modeling the kinetics of chemokine receptor binding, internalization and re-expression on alveolar macrophages (AM) as shown in **Figure 1**. The present agent-based simulation algorithm has linear time complexity in the number of agents and in the number of timesteps. Thus, treating single molecules as single virtual agents would render the simulations computationally intractable. Scalability in terms of constituent quantities is one of the strengths of partial differential equations (Horn et al., 2012) as the time complexity of our numerical method is linear in the number of grid points, molecule species and timesteps. In summary, treating cells at the microscopic level of discrete agents and molecules at the macroscopic level of continuous distributions ensures keeping the balance between computational tractability and detailed modeling across interwoven time- and length-scales (Guo et al., 2008). The source code of the hABM is available from the authors upon request.

2.2. Numerical Solution of the Reaction-Diffusion Equation in the Alveolus

2.2.1. Reaction-diffusion equation

The spatio-temporal distribution of chemokines on the inner surface of the alveolus is described by the following reaction-diffusion equation:

$$\frac{\partial c(\vec{r},t)}{\partial t} = D\Delta c(\vec{r},t) - \lambda\, c(\vec{r},t) + S(\vec{r},t) - Q(\vec{r},t). \qquad (1)$$

Here $c(\vec{r},t)$ denotes the molecular concentration of chemokines at position $\vec{r}$ and time t and Δ is the Laplace operator. The chemokine's isotropic diffusion coefficient is given by D and its degradation rate is given by λ. The spatio-temporal source of molecular concentration is represented by the term $S(\vec{r},t)$ associated with chemokine producing alveolar epithelial cells (AEC) of type I and type II. The term $Q(\vec{r},t)$ represents the uptake of chemokines by AM and is explained in detail below. Numerical integration of the reaction-diffusion Equation (1) within the surfactant on the inner alveolar surface requires a discretization of the thin fluidic lining layer by a grid with close-to-equidistant grid points.

2.2.2. Discretization of the Surfactant

Generating a grid with an arbitrary number of close-to-equidistant grid points on the surface of a spherical geometry is related to the Thomson problem (Thomson, 1904). This problem was raised more than a century ago in the context of finding the minimal electrostatic potential energy configuration for n equally charged particles that repel each other by Coulomb forces on the surface of a unit sphere. An equidistant distribution of points is beneficial for the numerical solution of the reaction-diffusion equation with regard to computing time and numerical stability. We take advantage of a crowd-based numerical approximation platform that determines the global minima using a variety of different optimization algorithms (MacWilliam and Cecka, 2013). Next, in order to obtain the neighborhood relationship between the grid points, we use the close-to regular distribution of points as inputs and compute the convex hull, where each of its edges corresponds to a pair of neighboring grid points. Note that the triangulation of discrete points on a sphere surface using the convex hull is equivalent to the Delaunay triangulation of these points in three dimensions (Brown, 1979). Finally, the dual

graph of the Delaunay triangulation, i.e., the Voronoi tesselation (De Berg et al., 2008), was computed in order to obtain the surface-area associated with each grid point, i.e., the area of the corresponding Voronoi cell. As will be shown below, this measure together with the length of the Voronoi edge between neighboring Voronoi cells are required for solving the reaction-diffusion Equation (1) numerically.

It should be noted that, since the human alveolus does not correspond to a full sphere, not each grid point belongs to the alveolar surface. In fact, each point of the grid can be labeled as one of the three categories: (i) alveolar surface point, (ii) boundary point, (iii) outside point. A point is considered to be an alveolar surface point if it is part of the alveolar three-quarter sphere and does not cover a pore of Kohn. All other points are outside points, except for boundary points which have at least one neighboring point being an alveolar surface point (see Supplementary Figure S1A and Video S1). We use absorbing boundary conditions in each simulation scenario, i.e., the concentration at each boundary point is kept fixed at zero for all times. The representation of the surfactant with an average thickness of only 200 nm (Bastacky and Lee, 1995) is based on 10^4 close-to-equidistant grid points of the spherical surface at an average distance of $4.45 \pm 0.16\,\mu$m (see Video S1). This allows resolving AEC of type I and type II that are, respectively, 60 μm and 9.3 μm in diameter, as well as the pores of Kohn that are 6 μm in diameter as estimated from literature data in Pollmächer and Figge (2014).

2.2.3. Numerical integration of the reaction-diffusion equation

The reaction-diffusion Equation (1) is numerically integrated in time using a finite difference method for unstructured grids as described by Sukumar (2003). Here, Voronoi cells are the placeholders of the chemokine concentrations, where each Voronoi cell may contain several molecular species. The kth Voronoi cell is associated with grid point $\vec{r}_k$ of the Delaunay triangulation and has area A_k and a finite set of neighbors $\mathcal{N}(k)$. The relation with neighboring Voronoi cells $\ell \in \mathcal{N}(k)$ is defined by the length of the Voronoi edge $h_{k\ell}$ and the Euclidean distance between the two Voronoi cells $d_{k\ell}$, as depicted in Supplementary Figure S1B. The numerical integration is then performed in a straightforward fashion over each Voronoi cell k that is associated with a grid point of the category alveolar surface point:

$$\tilde{c}(\vec{r}_k, t+\Delta t) = \tilde{c}(\vec{r}_k, t) + \Delta t\Big(\sum_{\ell \in \mathcal{N}(k)} D \frac{h_{k\ell}}{d_{k\ell} A_k} \big[\tilde{c}(\vec{r}_\ell, t) - \tilde{c}(\vec{r}_k, t)\big] - \lambda\, \tilde{c}(\vec{r}_k, t) + S(\vec{r}_k, t) - Q(\vec{r}_k, t)\Big)\,. \quad (2)$$

Here and in what follows the discretized concentration values are indicated by the symbol $\tilde{c}$. In our model, both AEC of type I and type II may secrete chemokines, which is appropriately captured by a non-vanishing source term $S(\vec{r}_k, t)$ at all grid points of the AEC associated with the conidium.

2.2.4. Validation of the numerical solution

In order to validate the implementation of the close-to-equidistant grid for the spherical system and the algorithm for the numerical solution of the reaction-diffusion Equation (1), we performed simulations of scenarios for which the analytical solutions are known. These scenarios were based on the analytical solution of the isotropic diffusion equation in terms of spherical harmonics (Sbalzarini et al., 2006). For a sphere with radius r and molecular diffusion coefficient D on its surface an analytical solution of the reaction-diffusion Equation (1) for vanishing molecule degradation and absent source- and reaction-term is given by

$$c(\vec{r} = (r, \vartheta, \varphi), t) = \sqrt{\frac{3}{4\pi}} \cos(\vartheta) \exp\left(-\frac{2D}{r^2} t\right)\,. \quad (3)$$

Here, surface positions $\vec{r}$ are represented using spherical coordinates with polar angle ϑ and azimuthal angle φ. Simulations were started from the initial condition $c(\vec{r}, t=0) = \sqrt{3/(4\pi)}\cos(\vartheta)$. The accuracy of the numerical solution was evaluated by comparing with the analytical solution on the spherical surface using biquadratic interpolation at 2×10^4 pre-defined close-to-equidistant points.

2.3. Chemotaxis Model of Alveolar Macrophages

The previously established agent-based model of the human alveolus (Pollmächer and Figge, 2014) is extended by modeling the interactions between molecule concentrations and chemokine receptors of AM, including the internalization of bound receptors and their subsequent re-expression on the AM surface. This enables AM to sense chemokine gradients that ultimately drive the migratory response of the phagocytes. Here, we essentially follow the receptor kinetics model as previously presented in Guo and Tay (2008) and Guo et al. (2008), apart from modifications required in the present context of modeling the dynamics of infection in the curved environment of a human three-quarter alveolus.

Since the average distance between neighboring grid points is four to five times smaller than the AM diameter of $r_{\mathrm{AM}} = 10.6\ \mu m$ (Krombach and Münzing, 1997), each AM is an agent associated with on average 20 grid points on the interactive molecule layer. In the reaction-diffusion Equation (1), the interaction between chemokines and AM receptors is represented by the term

$$Q(\vec{r}, t) = \sum_{m \in \mathcal{M}(t)} Q_m(\vec{r}, t)\,, \quad (4)$$

where $\mathcal{M}(t)$ is the set of AM present in the alveolus at time t. $Q_m(\vec{r}, t)$ denotes the reaction term of the mth AM with the chemokines in the surfactant, which is defined at each grid point q as follows:

$$Q_m(\vec{r}_q, t) = \begin{cases} \frac{k_b}{A_{\mathrm{AM}}}\, \tilde{c}(\vec{r}_q, t)\, [R]_m(t) & \text{, if } q \in \mathrm{cov}_m(t) \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $\mathrm{cov}_m(t)$ represents the set of covered grid points by the mth AM (see Supplementary Figure S2), $[R]_m(t)$ is the current

number of free receptors on the AM and k_b is the binding rate between AM receptors and the chemotactic cytokines in the surfactant. The interaction surface for the reaction between the receptors of the AM cell wall and the chemokines in the surfactant is denoted by $A_{\mathrm{AM}} = \pi r_{\mathrm{AM}}^2$. Beside the number of free receptors, each AM m is an agent keeping track of its current number of bound ($[LR]_m$) and internalized receptors ($[R^*]_m$). The kinetics of ligand-binding, receptor internalization and re-expression is described by a system of ordinary differential equations:

$$\frac{\mathrm{d}[R]_m(t)}{\mathrm{d}t} = k_r\,[R^*]_m(t) - A_{\mathrm{AM}} \sum_q Q_m(\vec{r}_q, t)\ , \quad (6)$$

$$\frac{\mathrm{d}[RL]_m(t)}{\mathrm{d}t} = A_{\mathrm{AM}} \sum_q Q_m(\vec{r}_q, t) - k_i\,[LR]_m(t)\ , \quad (7)$$

$$\frac{\mathrm{d}[R^*]_m(t)}{\mathrm{d}t} = k_i\,[LR]_m(t) - k_r\,[R^*]_m(t)\ . \quad (8)$$

Here, k_i is the internalization rate of bound receptors and k_r is the recycling rate associated with the re-expression of internalized receptors. All model parameters together with their experimentally relevant regimes of values are listed in **Table 1**. The parameters related to the receptor-kinetics model of AM, k_b, k_i and k_r, are fixed to the geometric means of their corresponding experimental range.

In our model, the kinetics of bound receptor differences along the current chemokine gradient is coupled to the directional persistence time of migrating AM (Farrell et al., 1990). Thus, as shown in Supplementary Figure S2, we consider that the mth AM weights the direction of the average chemokine concentration gradient $\vec{g}_m(t)$ in each timestep by the difference in newly bound receptors at the front and rear of the cell along the gradient.

The difference in the chemokine concentration across the interaction surface of the mth AM between its front and rear, $\Delta c_{m,\mathrm{diff}}$, is computed using the distance between the respective barycenters of the front und rear of this AM and its corresponding concentration gradient:

$$\Delta c_{m,\mathrm{diff}}(t) = \frac{8\,r_{\mathrm{AM}}}{3\pi} ||\vec{g}_m(t)||\ , \quad (9)$$

where the chemokine concentration gradient $\vec{g}_m(t)$ over the mth AM is obtained from averaging over the local gradients of all grid points covered by the mth AM ($\mathrm{cov}_m(t)$). Then, the difference in newly bound receptors between the front and rear of the AM along the current gradient per timestep Δt is

$$\Delta[LR]_{m,\mathrm{diff}}(t) = k_b\,\Delta c_{m,\mathrm{diff}}(t)\,\frac{[R]_m(t)}{2}\,\Delta t\ . \quad (10)$$

The most favorable direction of migrating AM is determined by computing the sum of weighted gradients over one period of directional persistence:

$$\vec{g}_{m,\mathrm{cum}}(t^*_{\mathrm{begin}}, t^*_{\mathrm{end}}) = \sum_{t=t^*_{\mathrm{begin}}}^{t^*_{\mathrm{end}}} \Delta[LR]_{m,\mathrm{diff}}(t)\frac{\vec{g}_m(t)}{||\vec{g}_m(t)||}\ , \quad (11)$$

where t^*_{begin} and t^*_{end} denote the start and the end time for the period of directional persistence.

Finally, after each period of directional persistence, the respective AM migrates in the direction of the weighted cumulative gradient $\vec{g}_{m,\mathrm{cum}}(t^*_{\mathrm{begin}}, t^*_{\mathrm{end}})$ with probability

$$p_{\mathrm{directed}} = \min(||\vec{g}_{m,\mathrm{cum}}(t^*_{\mathrm{begin}}, t^*_{\mathrm{end}})||\sigma_{\mathrm{AM}}\ ,\ 1)\ . \quad (12)$$

This probability is proportional to the bound receptor differences along the cumulative gradient (Devreotes and Zigmond, 1988) and the constant of proportionality is the AM sensitivity σ_{AM} that was determined by Tranquillo et al. (1988) (see **Table 1**).

2.4. System Setup for Simulation Studies

2.4.1. Steady state analysis

Initially, all grid points were set to zero molecular concentration and one permanently and homogenously secreting AEC of type I at the bottom of an otherwise empty three-quarter alveolus was placed. Keeping track of the time-dependent relative concentration change,

$$\Delta\tilde{c}_{\mathrm{rel}}(\vec{r}_k, t) \equiv \frac{\tilde{c}(\vec{r}_k, t+\Delta t) - \tilde{c}(\vec{r}_k, t)}{\tilde{c}(\vec{r}_k, t)}\ , \quad (13)$$

at grid points k, the steady state of the molecular distribution was considered to be reached when the maximum value of $\Delta\tilde{c}_{\mathrm{rel}}$ over all grid points fell below a threshold value of one permille. Measurements were repeated 50 times per parameter configuration and the results were averaged, keeping the number of randomly positioned pores of Kohn in the alveolus constant.

TABLE 1 | Parameters used for the chemotaxis model of alveolar macrophages.

Symbol	Description	Unit	Value	Experimental range	References
D	Chemokine diffusion coefficient (in water)	$\mu\mathrm{m}^2 \times \mathrm{min}^{-1}$	Varied	$6 \times 10^2 - 3.5 \times 10^4$	Francis and Palsson (1997); Randolph et al. (2005)
s_{AEC}	Chemokine secretion rate of AEC	min^{-1}	Varied	Unknown	
λ	Chemokine degradation rate	min^{-1}	Varied	$3 \times 10^{-3} - 4.2 \times 10^{-2}$	Beyer and Meyer-Hermann (2008)
k_b	Ligand-receptor binding rate	$\mu\mathrm{m}^2 \times \mathrm{min}^{-1}$	1×10^{-2}	$7 \times 10^{-4} - 0.3$	Sklar (1984); Pelletier (2000); Guo et al. (2008)
k_i	Receptor internalisation rate	min^{-1}	7×10^{-2}	$3 \times 10^{-3} - 1.8$	Beyer and Meyer-Hermann (2008); Guo et al. (2008)
k_r	Receptor recycling rate	min^{-1}	5×10^{-2}	$6 \times 10^{-3} - 0.5$	Beyer and Meyer-Hermann (2008); Guo et al. (2008)
R_0	Initial number of chemokine receptors		5×10^4	$2 \times 10^4 - 2 \times 10^5$	Beyer and Meyer-Hermann (2008); Guo et al. (2008)
σ_{AM}	Sensitivity to bound-receptor differences		1.2×10^{-3}	1.2×10^{-3}	Devreotes and Zigmond (1988); Farrell et al. (1990)

2.4.2. Virtual infection scenario

For studying *A. fumigatus* infection in a three-quarter alveolus with constant radius, the virtual infection scenario from Pollmächer and Figge (2014) was followed. At $t = 0$ a binomially distributed number of AM and the conidium were placed randomly over the surface of the three-quarter alveolus and all grid points were set to zero molecular concentration. AM migrated according to a biased persistent random walk with constant speed v and constant directional persistence time t_p and were able to leave or enter the alveolus at either a pore of Kohn or the alveolar entrance ring. The position of the conidium was fixed over the whole simulation and migration of AM followed the chemotaxis model that was here previously introduced. In each virtual infection scenario the AEC of type I or II that was associated with the randomly positioned conidium released the chemoattractant permanently and homogenously with a constant secretion rate s_{AEC}. The simulation ended at the first physical contact between an arbitrary AM and the conidium. The diffusion coefficient D of the chemokine was varied over a wide range in order to account for the viscosity of the surfactant that is expected to be higher than that of water and to which experimental ranges are typically referring. In **Table 1**, the parameter regimes of the chemotaxis model are summarized and the values that were varied in the simulations are indicated.

2.4.3. Virtual infection scenario including gradient-based recruitment of alveolar macrophages

In Pollmächer and Figge (2014), AM insertion into the three-quarter alveolus followed a uniform distribution over the length of the boundary line elements. Numerical values of chemokine concentrations allow for recruitment of AM from neighboring alveoli based on the strength of the gradient. Realization of gradient-based recruitment was implemented in the following way. First, on AM entrance into the alveolus the maximum gradient was computed, $\max\{||\vec{g}_b(t)||\}$, over the finite set of edges of the triangulated grid that cross the boundary. The pairs of vertices corresponding to these edges each held one vertex labeled as boundary point and the other one labeled as alveolar surface point. Secondly, a uniformly distributed random boundary point $\vec{r}_{b,random}$ was drawn from all possible boundary points and the corresponding probability of AM insertion was calculated as follows:

$$p_{in}(\vec{r}_{b,random}, t) = \frac{||\vec{g}(\vec{r}_{b,random}, t)||}{\max\{||\vec{g}_b(t)||\}}. \quad (14)$$

This probability was used for stochastic AM insertion at position $\vec{r}_{b,random}$ and was realized by a Monte Carlo acceptance-rejection method to sample the gradient-based probability distribution of AM insertion over the boundary points. On rejection of a boundary point a new one was drawn with probability $p_{in}(\vec{r}_{b,random}, t)$ followed by another Monte Carlo decision until a boundary point was accepted. As before, first-passage-time simulations were performed over 10^3 repetitions for each parameter configuration.

3. Results

3.1. Hybrid Agent-Based Model Reproduces Analytical Solutions

We evaluated and validated the numerical solution of our PDE solver by comparison with an analytical solution over the surface of a full sphere (see Section 2.2.4 for details). The mean of the absolute and relative errors per timestep were computed for both varying timesteps and varying numbers of grid points in order to demonstrate the accuracy of the numerical method (see **Figure 2**). The method shows first-order accuracy in the timestep as the absolute and relative mean errors per timestep scale quadratically. Furthermore, it is observed that numerical instability occurs for too large values of Δt, as is expected for an explicit forward-Euler approach. To guarantee

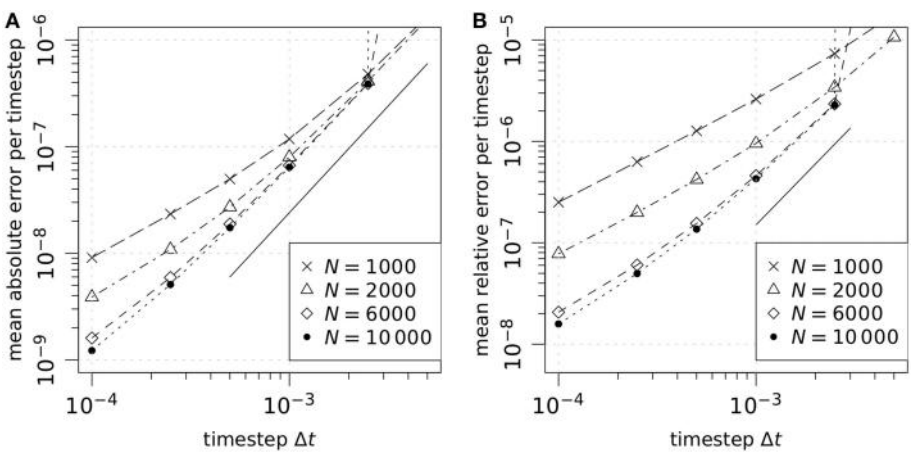


FIGURE 2 | Numerical error analysis of the PDE solver for Equation (2) on the spherical surface with $\lambda = 0$, $S(\vec{r}_k, t) = 0$, and $Q(\vec{r}_k, t) = 0$ at each grid point k. Simulations were carried out in an alveolus with a radius $r = 116.5\,\mu$m from time $t = 0$ min to a final time $t = 1$ min with an isotropic diffusion coefficient of $D = 2000\,\mu m^2$/min. The mean absolute error **(A)** and mean relative error **(B)** per timestep of our PDE solver for different numbers of grid points N and timesteps Δt are compared to the theoretically expected quadratic scaling (*solid line*).

numerical stability in our simulations, we determined the limits of numerical stability for different diffusion coefficients D over the set of grid points $\mathcal{G}$ using the condition

$$\Delta t \leq \min_{k \in \mathcal{G}} \left[\left(D \sum_{l \in \mathcal{N}(k)} \frac{h_{kl}}{d_{kl} A_k} \right)^{-1} \right] \tag{15}$$

and adjusted the global simulation timestep Δt one order of magnitude lower than the respective limits.

3.2. Steady State of Alveolar Chemokine Distribution Reached within Hours

We performed a numerical study to characterize the steady state of the alveolar chemokine distribution in terms of the concentration profile and the time required to reach the steady state (see Video S2 for the transition from the onset of AEC secretion into steady state). Simulations were carried out using one permanently and homogenously secreting AEC of type I in the bottom of an empty alveolus (see Section 2.4.1 for details) In **Figure 3**, we summarize the results of the steady state analysis for varied diffusion coefficients, degradation rates and secretion rates of the chemokine. Interestingly, we found that the time required to reach the steady state depends on the values for the diffusion coefficient D and the degradation rate λ but not on the amount of chemokine secretion s_{AEC} per time (**Figure 3A**). In the absence of degradation, the time required to reach the steady state ranges from 4 min for a diffusion coefficient of $D = 6000\,\mu\mathrm{m}^2/\mathrm{min}$ to 8.5 h for a diffusion coefficient of $D = 20\,\mu\mathrm{m}^2/\mathrm{min}$. In the presence of degradation, the times required to reach the steady

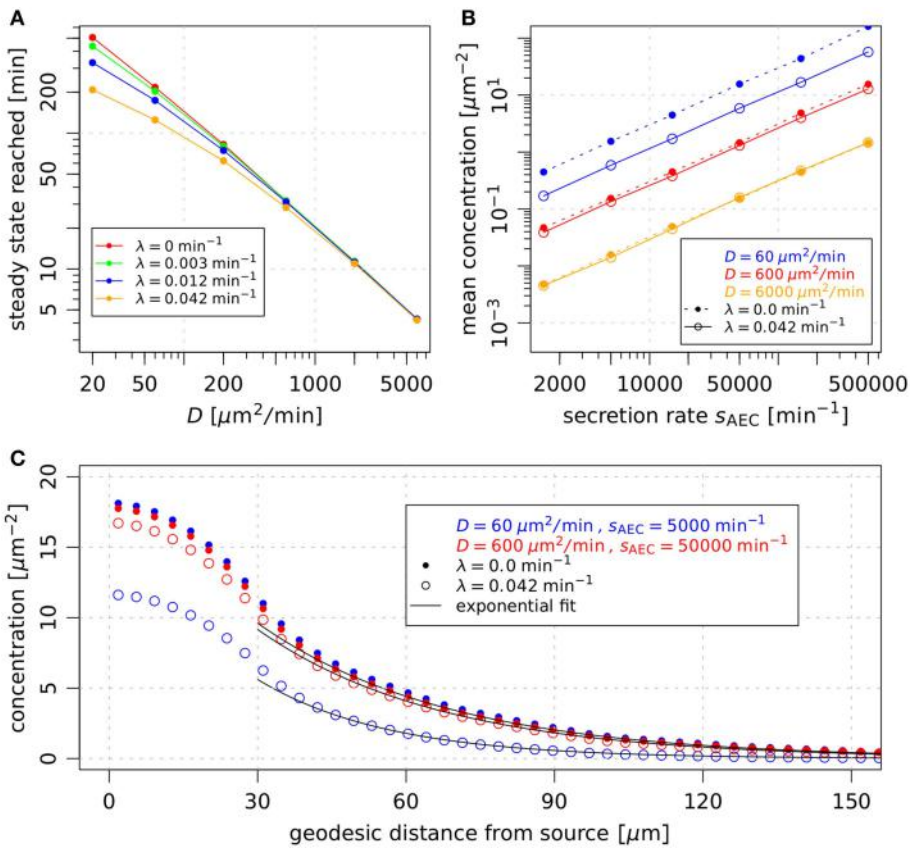


FIGURE 3 | Steady state analysis of the concentration profile in the alveolus for varied diffusion coefficients *D*, secretion rates s_{AEC} and degradation rate λ. One permanently and homogenously secreting source with radius $r_{\mathrm{AEC}} = 30\,\mu$m was placed in the bottom of the three-quarter alveolus and tthe relative concentration changes $\Delta\tilde{c}_{\mathrm{rel}}$ (see Equation 13) were tracked over time at each grid point k. The steady state of the molecular distribution was considered to be reached when the maximum value of $\Delta\tilde{c}_{\mathrm{rel}}$ over all grid points fell below a threshold value of 0.001. **(A)** Comparison of the mean values of the time when steady state was reached for different degradation rates and diffusion coefficients averaging over the secretion rates $\{1.5 \times 10^3, 5 \times 10^3, 1.5 \times 10^4, 5 \times 10^4, 1.5 \times 10^5, 5 \times 10^5\}\,\mathrm{min}^{-1}$. Each mean value has a relative standard deviation less than five percent. **(B)** Average concentration over all grid points labeled as alveolar surface point at steady state. **(C)** Concentration profile at steady state as a function of the geodesic distance from the center of the source. In each simulation concentration values were averaged over points of the three-quarter sphere with equivalent geodesic distance from the center of the source. Here biquadratic interpolation was used to obtain the concentration value at arbitrary points on the alveolar surface. Afterwards the means over simulation runs with identical parameter configuration were computed. We applied exponential fits to each concentration profile using least squares to optimize the parameters a and b in the function $c(x) = a\exp(bx)$ over concentration values at geodesic distances above the AEC radius $r_{\mathrm{AEC}} = 30\,\mu$m.

state were systematically decreasing with increasing degradation rates in a diffusion-dependent fashion (**Figure 3A**).

In **Figure 3B** it can be seen that the parameter variation lead to average concentration values that span a range of five orders of magnitude from $10^{-2}\,\mu\mathrm{m}^{-2}$ to $10^{2}\,\mu\mathrm{m}^{-2}$. The mean concentration was observed to increase linearly with increasing secretion rate s_{AEC}. We found that different parameter combinations showed similar mean concentration values and almost identical concentration profiles over the geodesic distance from the secreting AEC (see **Figure 3C**). Irrespective of the secretion rate, diffusion coefficient and degradation rate the profile of concentration over the surface of the alveolus showed an exponential distance-dependence from the secreting AEC for geodesic distances larger than the radius of the secreting AEC.

We generally observed that the impact of chemokine degradation on the time to reach the steady state and on the amount and profile of the chemokine concentration is largest for small diffusion coefficients (see **Figures 3A–C**). This is a direct consequence of reduced molecule motion, because on average molecules remain in the alveolus for a longer time period before leaving through a pore of Kohn or through the alveolar entrance ring. In particular, the time required to reach the steady state, the average chemokine concentration as well as the level of the concentration profile were lowered for elevated degradation rates. These effects were depending on the diffusion coefficient: While for diffusion coefficents $D \geq 2000\,\mu\mathrm{m}^2/\mathrm{min}$ all three observables were reduced by less than 5% relative to the case with absent degradation, for $D \leq 60\,\mu\mathrm{m}^2/\mathrm{min}$ this reduction was observed to increase up to 85%. For example, in the extreme case of the small diffusion coefficient $D = 20\,\mu\mathrm{m}^2/\mathrm{min}$ and at a secretion rate of 1.5×10^4 molecules per minute, the average concentration ranges between 2.3 and 14.7 molecules per $\mu\mathrm{m}^2$ and the time required to reach the steady state varied in a degradation-dependent fashion between 3.5 and 8.5 h.

3.3. Virtual Infection Model Reveals Relevant Parameter Regimes

We performed computer simulations on the early immune response against *A. fumigatus* infection mediated by chemokines that are released from the AEC associated with the conidium. In contrast to our previous study, where chemotaxis was modeled in a simplified fashion by a probabilistic rule (Pollmächer and Figge, 2014), we here implemented a numerical solver for the reaction-diffusion equation extending over the inner surface of the alveolus. Thus, in the present implementation AM performed a biased persistent random walk and the directional bias was derived from local sensing of the current chemokine gradient by AM. The relative impact of directional over random migration was inferred from the difference in newly bound AM receptors along the gradient. Computer simulations with the refined AM chemotaxis model, which is described in the Section 2 and depicted in Supplementary Figure S2, enabled us to narrow down the regime of relevant parameters in terms of the diffusion coefficient, the degradation rate and the secretion rate of the postulated chemokine.

3.3.1. First-passage-times are mainly determined by diffusion coefficients and secretion rates

We measured first-passage-times in the alveolus, i.e., the time of first contact between AM and the conidium (see Video S3), in order to determine the requirements on the chemokine properties for a successful discovery of the fungal pathogen before the onset of germination (see Section 2.4.2 for details). First-passage-times were computed for 864 different parameter combinations (see Supplementary data in Supplementary Material) and for each combination 10^3 simulations of the *A. fumigatus* infection scenario were performed to obtain statistically firm results. From the distributions of first-passage-times, we computed the fraction of first-passage-times above 6 h, $p(\mathrm{FPT} > 6\,\mathrm{h})$, where 6 h were chosen based on the typical time period required for *A. fumigatus* germination. The results are presented in **Figure 4** and demonstrate, in agreement with our previous study (Pollmächer and Figge, 2014), that AM with migration speed $v = 2\,\mu\mathrm{m/min}$ exceeded the first-passage-time of 6 h in more than 5 % of the simulations for all parameter combinations (see short-dashed lines in **Figures 4A,D,G**). A comparison of **Figures 4A,D** shows that an increase in the persistence time from $t_{\mathrm{p}} = 1\,\mathrm{min}$ to $t_{\mathrm{p}} = 2\,\mathrm{min}$ was always associated with a decrease of $p(\mathrm{FPT} > 6\,\mathrm{h})$. Next, we found that taking molecular degradation into account did not have a strong impact on $p(\mathrm{FPT} > 6\,\mathrm{h})$, as can be observed by comparing **Figures 4D,G** for $t_{\mathrm{p}} = 2\,\mathrm{min}$. These observations remain qualitatively the same for higher migration speeds of AM, see **Figures 4B,E,H** for $v = 4\,\mu\mathrm{m/min}$ and **Figures 4C,F,I** for $v = 6\,\mu\mathrm{m/min}$. However, higher migration speeds of AM do have a quantitative impact on $p(\mathrm{FPT} > 6\,\mathrm{h})$.

The dashed-dotted and long-dashed lines in **Figure 4** indicate the values of $p(\mathrm{FPT} > 6\,\mathrm{h})$ for AM performing, respectively, a persistent random walk and a biased persistent random walk, as previously simulated in Pollmächer and Figge (2014). The persistent random walk of AM always marks an upper limit for $p(\mathrm{FPT} > 6\,\mathrm{h})$, i.e., first-passage-times are on average always decreased in the presence of chemotaxis, as could be expected for a low concentration of chemokines in the alveolus. On the other hand, compared to the biased persistent random walk model the performance of the chemotaxis model could yield lower values for $p(\mathrm{FPT} > 6\,\mathrm{h})$, depending on the combination of parameters. In particular, we found that this is the case for combinations of a relatively high secretion rate and a relatively low diffusion constant. Note that the probabilistic rule for biased persistent random walk as previously simulated in Pollmächer and Figge (2014) was coupled to the direction of the shortest path from the AM to the AEC associated with the conidium. Occasionally, AM could leave the alveolus through a pore of Kohn if one of them was along the respective path of migration. In the present approach the frequency of this event was reduced, due to preferred AM migration in the direction of the chemokine gradient, which generally pointed away from pores of Kohn (see Videos S2 and S3). In summary, the diffusion coefficient and the secretion rate were again found to be the most important parameters, whereas the value of the degradation rate had only minor impact on $p(\mathrm{FPT} > 6\,\mathrm{h})$ (see **Figures 4G–I**).

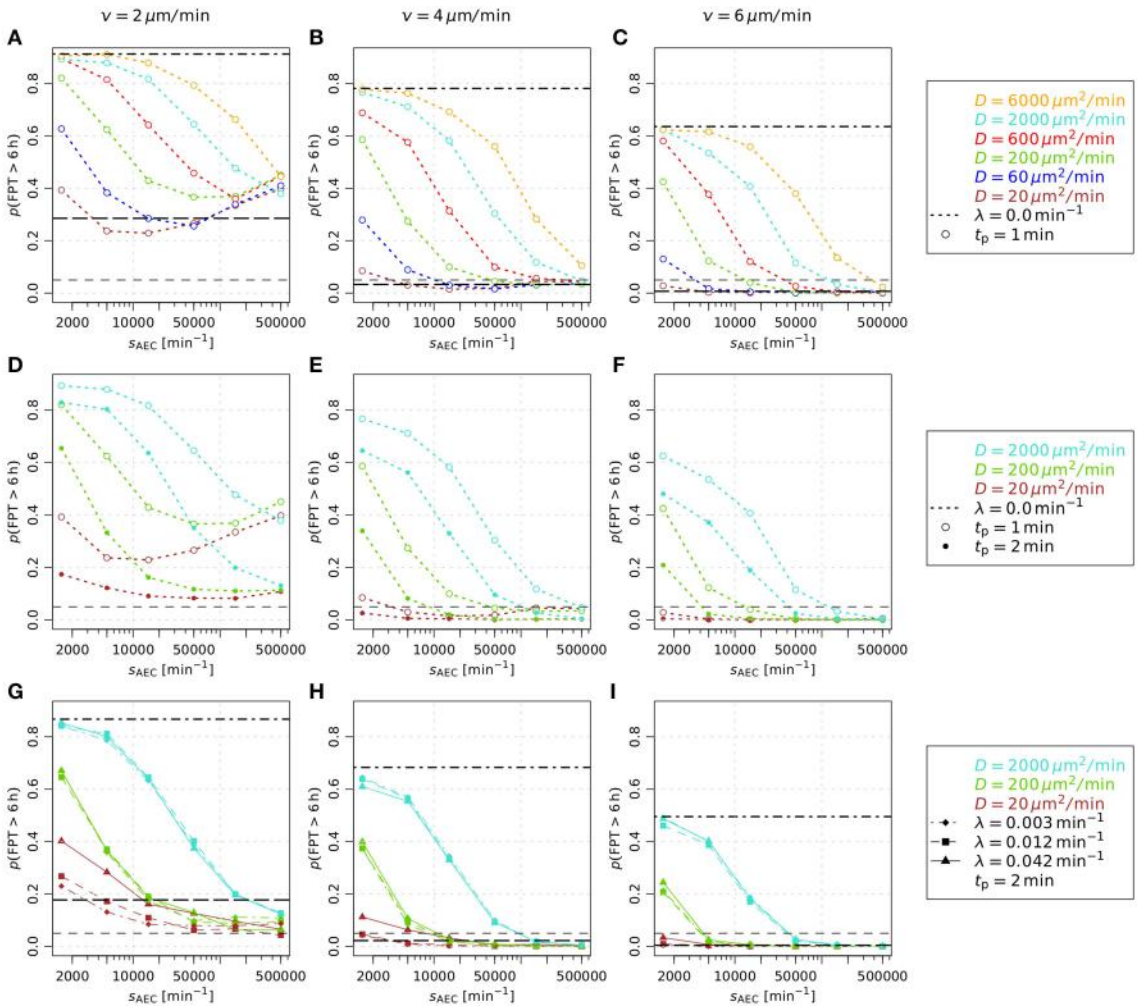


FIGURE 4 | Analysis of first-passage-time distributions varying the macrophage related parameters migration speed and persistence time and varying the chemokine related parameters diffusion coefficient, secretion rate and degradation rate. In each subfigure **(A–I)** the fraction of first-passage-times above 6 h, $p(\text{FPT} > 6\,\text{h})$, is plotted against the secretion rate of the AEC associated with the fungal conidi. The calculation of this fraction is based on the first-passage-time distribution which was derived performing 1000 first-passage-time simulations per parameter configuration. The results of the present study are compared to the biased persistent random walk (*long-dashed black line*) and the persistent random walk (*dashed-dotted black line*) by Pollmächer and Figge (2014). The *short-dashed black line* denotes the threshold $p(\text{FPT} > 6\,\text{h}) = 0.05$. In **(A–C)** the focus is on the variation of diffusion coefficients, **(D–F)** show the results for different persistence times t_p and **(G–I)** demonstrate the influence of the degradation rate λ.

Interestingly, we observed a minimum of $p(\text{FPT} > 6\,\text{h})$ as a function of the secretion rate for various diffusion coefficients in the case of AM migration speed $v = 2\,\mu\text{m/min}$ and persistence time $t_\text{p} = 1\,\text{min}$ (see **Figure 4A**). This system behavior reflects the fact that an optimal concentration of chemokines exists for an efficient guidance of AM. The value of the optimal concentration is determined by the interplay of several factors, e.g., the secretion rate, diffusion coefficient and degradation rate of the chemokine as well as the number of AM receptors and their dynamics of binding, internalization and re-expression. For example, a too high chemokine concentration is associated with a low number of unbound AM receptors limiting the adaptation of AM migration along the chemokine gradient. We further analyzed this situation by computing the probability of directed AM migration for different secretion rates and for AM migration speeds $v = 2\,\mu\text{m/min}$ and $v = 4\,\mu\text{m/min}$. The resulting probability distributions are shown in **Figure 5** as a function of the geodesic distance of AM from the AEC associated with the conidium. We found that optimal values of $p(\text{FPT} > 6\,\text{h})$ in **Figure 4A** correspond to probability distributions with a narrow and peaked maximum (see red curves in **Figure 5**). For a constant diffusion coefficient, lower secretion rates were associated with less prominent maxima in the probability distribution (see blue curves in **Figure 5**), which in turn increased $p(\text{FPT} > 6\,\text{h})$. On the other hand, higher secretion rates were associated with extended and flat maxima at relatively large geodesic distances

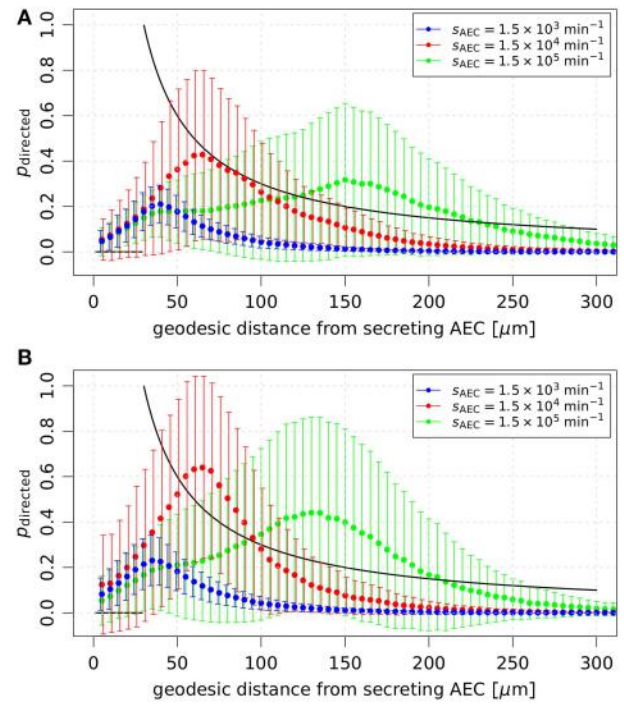


FIGURE 5 | Probabilities of directed AM migration over the geodesic distance from the AEC associated with the conidium. The mean and standard deviation of the probability $p_{directed}$ are shown in the absence of chemokine degradation for the diffusion coefficient $D = 60\,\mu m^2$/min with AM directional persistence time $t_p = 1$ min. In **(A)** AM migrate with speed $v = 2\,\mu$m/min and in **(B)** with speed $v = 4\,\mu$m/min. Averages and standard deviations were determined using the probabilities of directed AM migration that were drawn over the whole simulation time in all simulation runs. The present results are compared to the probabilistic rule for directed migration (*solid black line*) used in Pollmächer and Figge (2014).

from the boundary of the secreting AEC (see green curves in **Figure 5**). It should be noted that the profiles of the determined probability distributions are the results of various factors, such as the chemokine concentration and the receptor dynamics of AM. For example, in the case of high secretion rates, many AM receptors were already bound to the chemokine at early time points due to its relatively high concentration in the alveolus. As a result, AM were guided to the AEC associated with the conidium relatively early in time. However, the relatively high concentration of chemokines also had the adverse effect that the number of free AM receptors was decreased at distances close to the secreting AEC. Consequently, fewer events of receptor-ligand binding lead to relatively low probabilities for directed AM migration and ultimately increased $p(\text{FPT} > 6\,\text{h})$.

An overview of the relevant combinations of model parameters for successful detection of the *A. fumigatus* conidium by AM is given in **Figure 6** for AM migration speed $v = 4\,\mu$m/min (A) and $v = 6\,\mu$m/min (B). As in Pollmächer and Figge (2014), we considered a parameter combination to be successful, if the value of $p(\text{FPT} > 6\,\text{h})$ was below five percent. Interestingly, the ratio between the secretion rate and the diffusion coefficient, s_{AEC}/D, was found to subdivide the parameter space into regimes of successful and unsuccessful parameter combinations. For $v = 4\,\mu$m/min and $t_p = 1$ min, successful detection occurred for $s_{AEC}/D \geq 250\,\mu m^{-2}$ (see **Figure 6A**). Moreover, with increasing directional persistence time and/or migration speed of AM this threshold was found to be systematically reduced. While the combinations $(v, t_p) = (4\,\mu\text{m/min}, 2\,\text{min})$ and $(v, t_p) = (6\,\mu\text{m/min}, 1\,\text{min})$ both shared the condition $s_{AEC}/D \geq 75\,\mu m^{-2}$, for $(v, t_p) = (6\,\mu\text{m/min}, 2\,\text{min})$ this threshold s_{AEC}/D was lowered to the value $25\,\mu m^{-2}$. To summarize, we found that the successful detection of the conidium by AM required the ratio between the secretion rate and the diffusion coefficient to be above a specific threshold, whereas the degradation rate had only minor impact on the first-passage-time (see **Figures 4**, **6**).

3.3.2. Gradient-based recruitment of AM increases relevant parameter regimes

Next, we studied a modification of AM insertion into the system at the boundaries, i.e., the alveolar entrance ring and the pores of Kohn. Previously, AM entered the three-quarter alveolus following a uniform random distribution over the length of the line elements belonging to all boundaries (Pollmächer and Figge, 2014). In the modified setup, we accounted for the time-evolution of the chemokine gradients at the boundaries by specifying probabilities for AM insertion according to the respective gradient strengths. In other words, AM insertion is more likely at boundaries with higher chemokine gradients (see Section 2.4.3 for details).

In **Figures 6C–E** it can be seen that gradient-based recruitment of AM generally increased the regime of parameter combinations for successful detection. At AM speeds of $4\,\mu$m/min and $6\,\mu$m/min the ratio of secretion rate to diffusion coefficient was systematically reduced (see **Figures 6D,E**). In contrast to the case where AM insertion was not gradient-based, in the present case a successful detection was also achieved at the migration speed of $2\,\mu$m/min for specific parameter combinations (see **Figure 6C**). Interestingly, for the AM parameters $(v, t_p) = (2\,\mu\text{m/min}, 1\,\text{min})$ the subdivision of the parameter space into regimes of successful and unsuccessful parameter combinations was not only determined by the ratio s_{AEC}/D. We checked that $p(\text{FPT} > 6\,\text{h})$ has a dependence on the secretion rate similar to the simulations in the absence of gradient-based AM insertion (see **Figure 4A**). However, in the present case the minimum of $p(\text{FPT} > 6\,\text{h})$ reached values below the five percent threshold for a limited range of the secretion rates that gave rise to the triangular region (see **Figure 6C**, blue area). The virtual infection model with gradient-based recruitment underlines the importance of chemokine-induced AM insertion points relative to the conidium position, as the results display a beneficial effect for the immune response of the host.

4. Discussion

In this study, we implemented a hybrid agent-based model (hABM) for *A. fumigatus* infection in human alveoli under physiological conditions to decipher the properties of a chemoattractant responsible for guiding alveolar macrophages

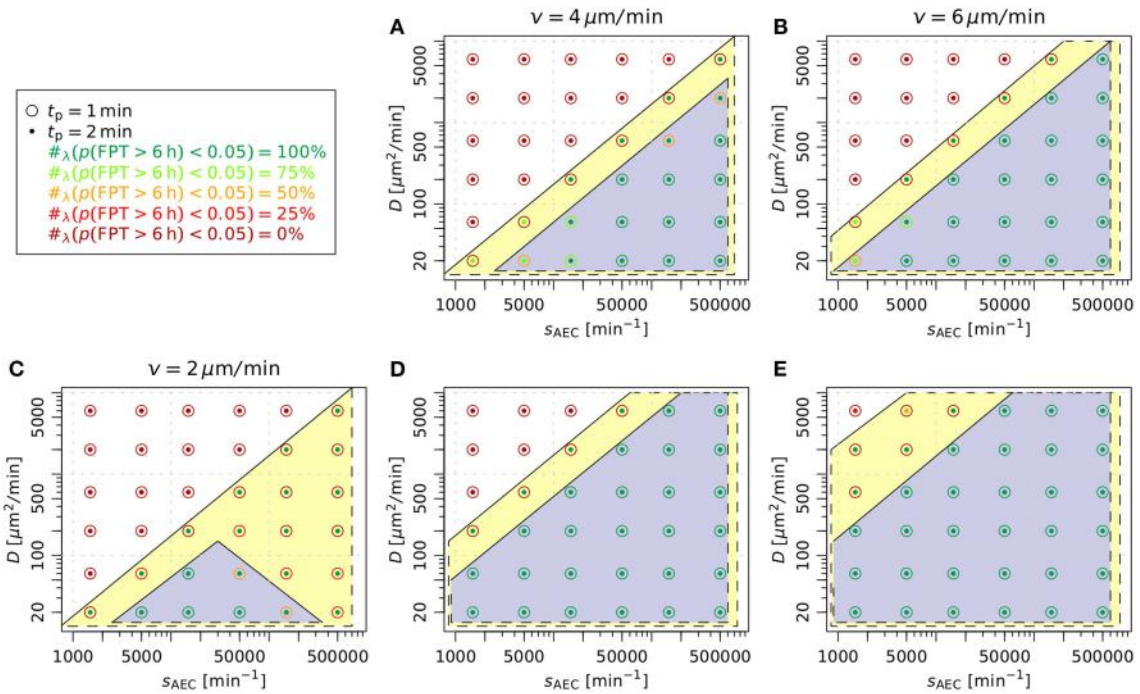


FIGURE 6 | Evaluation of parameters related to the AM chemoattractant based on the fraction of first-passage-times above 6 h. (A,B) summarize the results from **Figure 4**, where AM insertion followed a uniform random distribution over the length of the boundaries of the three-quarter alveolus. The case $v = 2\,\mu$m/min was left out as all parameter combinations lead to $p(\text{FPT} > 6\,\text{h}) \geq 0.05$. In **(C–E)** AM were inserted into the alveolus following a gradient-based probability distribution over the line elements belonging to the boundaries. The highlighted areas denote the regimes of parameters leading to timely detection of the conidium for the directional persistence times $t_p = 1$ min (*blue area*) and $t_p = 2$ min (*yellow area*) of AM under different migration speeds v. The variable $\#_\lambda(p(\text{FPT} > 6\,\text{h}) < 0.05)$ denotes the fraction of simulated degradation rates that lead to $p(\text{FPT} > 6\,\text{h}) < 0.05$ for a specific combination of parameters D and s_{AEC}.

(AM). The multi-scale simulations account for the dynamics at the cellular and molecular level, as well as the kinetics of binding, internalization and re-expression of chemokine receptors on AM. To scan the parameter space for combinations of parameters that ensure the timely detection of a conidium in the alveolus, we performed more than a million simulations of virtual infection scenarios in the experimentally relevant regimes. We were able to show that successful detection of the pathogen by AM is governed by the choice of five experimentally undetermined parameters: migration speed v and directional persistence time t_p of AM as well as the secretion rate s_{AEC}, diffusion coefficient D and the degradation rate λ of the chemokine.

Simulations of the chemokine dynamics on the inner surface of the alveolus with its peculiar boundary conditions were performed using an efficient and accurate finite difference method on Voronoi cells (Sukumar, 2003) to solve the reaction-diffusion equation on an unstructured triangular Delaunay grid with close-to-equidistant grid points. We first studied the chemokine profile in steady state under varying conditions in an empty three-quarter spherical alveolus. Our results show that, depending on the diffusion coefficient of the chemokine, the time until a steady state is reached can vary from several minutes for $D \geq 2000\ \mu$m^2/min to several hours for $D \leq 60\ \mu$m^2/min. This revealed that our previous study, where the chemokine dynamics was simplified by a probabilistic rule, is limited to infection scenarios in the limit of high diffusion coefficients (Pollmächer and Figge, 2014). In contrast, using the present approach we are in the position to study *A. fumigatus* infection from the onset of chemokine secretion by alveolar epithelial cells (AEC) induced by the conidium and extending over the time period of establishing a concentration profile until the conidium is successfully found by one of the AM.

Since it was shown that AM require chemotactic cues in order to timely detect the conidium before the start of germination (Pollmächer and Figge, 2014), we here developed the hABM to account for the spatio-temporal concentration of chemokines in the alveolus. We implemented the receptor-kinetics chemotaxis model of Guo et al. (2008) for AM migration on a grid with high spatial resolution to capture the spherical alveolar surface with the pores of Kohn. The chemotaxis model accounts for the binding of G protein-coupled receptors on the surface of AM to the AEC-derived chemoattracting ligands in the alveolar lining layer (surfactant). In general, eukaryotic cells are able to sense spatial differences in receptor occupation along the chemokine gradient by their relatively large size of at least 10 μm (van Haastert and Postma, 2007). In order to sense shallow gradients in the chemokine concentration of 1–5 %, chemotactic cells are in addition able to sense temporal differences in receptor occupation, which increases the signal-to-noise ratio and implies higher probabilities of polarization

directed along the gradient (van Haastert and Postma, 2007). We extended the chemotaxis model of Guo et al. (2008) by implementing AM sensing of the cumulated number of newly bound receptors over directional persistence times. This approach advances our previously applied phenomenological chemotaxis model, which was based on a constant function for the distance-dependent gradient strength (Pollmächer and Figge, 2014). In the present study, AM were enabled to sense dynamically changing local chemokine gradient strengths, which implicitly contained morphological information, e.g., concentration gradients pointing away from boundary elements. Simulations of the virtual infection scenario indicated that the present AM chemotaxis model unifies the random migration and chemotactic migration modes of our previous study in one model. In particular, we showed that persistent random walk was resembled for relatively low chemokine concentrations in the alveolus.

The computation of the first-passage-time, i.e., the duration until the conidium is detected by an AM for the first time, revealed the relative importance of the parameters associated with the chemokine distribution: the diffusion coefficient and the rate of chemokine secretion by the AEC associated with the conidium turned out to have a major impact, whereas chemokine degradation played a minor role. In particular, we found that the AEC secretion rate and the diffusion coefficient had counteractive effects regarding the average concentration of chemokines in the surfactant, i.e., decreasing the secretion rate lowered the average concentration whereas decreasing the diffusion coefficient increased it. Chemokines are diffusing in the alveolar lining layer (surfactant), which shields AM from the alveolar airspace, reduces surface tension and provides immunoregulatory proteins (Herzog et al., 2008; Hasenberg et al., 2013). In comparison with chemokine diffusion in water, the relatively high viscosity of the surfactant (Alonso et al., 2005) has the crucial effect to reduce the diffusivity of chemokines and by that to lower the AEC secretion rate required for the timely detection of the pathogen. We found the ratio of the AEC secretion rate to the diffusion coefficient, s_{AEC}/D, to be the main indicator for the outcome of the infection scenario. For specific values of the AM migration speed and directional persistence time, this ratio subdivided the parameter space into regimes of successful and unsuccessful parameter combinations, whereas this separation was only weakly depending on relevant degradation rates. The degradation rate showed to have some impact in virtual infection scenarios with relatively low diffusion coefficients, which was also the case in the simulations associated with the steady state analysis. Thus, decisive reduction of the chemokine amount available to AM due to molecular degradation is only of importance for a highly viscous surfactant. The specific morphology of human alveoli plays an important role in this regard as chemokine reduction was also a consequence of chemokine absorption at the pores of Kohn and the alveolar entrance ring. A relative dominance of chemokine decrease due to alveolar boundaries was determined for relatively high diffusion coefficients, whereas relatively low diffusion coefficients were accompanied with relatively high chemokine degradation. This was attributed to reduced molecule motion for reduced diffusion coefficients, thus, on average molecules remained in the alveolus for a longer time period before leaving through the alveolar boundaries. As observed in our previous study (Pollmächer and Figge, 2014), AM required a minimal migration speed of at least $4\,\mu$m/min to discover the fungal conidium before the onset of germination. However, as shown in the present study, assuming a recruitment of AM from neighboring alveoli that was based on the local chemokine gradient, an average speed of $2\,\mu$m/min was as well successful for a specific subset of parameter combinations. This finding is particularly interesting, because the actual AM migration speed in the alveolus is not known today, but is typically expected to be low (Hasenberg et al., 2013). Generally, our results show that the communication between different types of host immune cells and their reaction to threatening invaders needs to be finely tuned in order to mount and orchestrate a fast and adequate response.

The specific chemokine and AM receptor that are involved in the directed migration are not known today. It is well-known that AM express, for example, the chemokine receptor CXCR2 (Miller et al., 2003) that binds to the cytokine IL-8. Moreover, the presence of complement proteins in the surfactant yields the cleavage product C5a, and this anaphylatoxin is a potential candidate for which AM chemoattraction was observed (Farrell et al., 1990; Zipfel and Skerka, 2009). Resting conidia of *A. fumigatus* activate the complement system entirely by the alternative pathway (Kozel et al., 1989). Upon activation, C3 is cleaved into C3b and C3a, with C3b opsonizing the fungal surface and increasing uptake rates by macrophages (van Lookeren Campagne et al., 2007). Furthermore, C3b induces cleavage of C5 which leads to the production of the prominent proinflammatory and chemoattracting cytokine C5a (Brakhage et al., 2010). However, it is also known that resting *A. fumigatus* conidia reduce the impact of the complement cascade by binding complement regulatory proteins—such as factor H, FHL-1, CFHR-1, C4BP and plasminogon—and by that reducing the deposition of C3b molecules on their surface (Behnsen and Hartmann, 2008). These data suggest that single conidia do both trigger and counteract the complement cascade, such that the mediated stimulus of chemoattraction and inflammation is relatively weak and spatially confined. Nevertheless, it is conceivable that these signals can be detected by the AEC associated with the conidium and that this cell responds with the secretion of the chemokines for AM recruitment. Supporting evidence for this hypothesis is provided by a study of rat AEC of type II: binding of C5a to these cells lead to increased expression of the C5a receptor on the AEC surface and to the production of macrophage inflammatory protein-2 as well as neutrophil-chemoattractant-1 (Riedemann et al., 2002).

Our computational approach to investigate *A. fumigatus* infection complements wet lab experiments. *In vivo* measurements suffer from the circumstance that they can only be carried out with high doses of conidia that do not reflect the physiological condition of daily inhalation rates of a few thousand conidia (O'Gorman and Fuller, 2008; Pollmächer and Figge, 2014). The agent-based modeling approach allows studying the early immune response, i.e., we modeled a setting with those immune cells that are resident in alveoli and

performed virtual infection simulations to low numbers of conidia in a physiologically reasonable host-setting. Simulations enabled narrowing down the experimentally relevant regime of parameters to a subset of potential parameter combinations for healthy individuals. These predictions may initiate further wet lab investigations that should focus on quantitative aspects of the early immune response, e.g., the relative contributions of the complement system and the alveolar epithelial cells to the daily challenge with *A. fumigatus* or the identification of the specific chemokine for AM and the rate at which it is secreted by AEC. Furthermore, if possible by sophisticated imaging techniques in the future, it will be highly interesting to determine values of AM migration speed and migration mode in their natural environment to clarify their general role in the immune response, e.g., as compared to neutrophil migration in the alveolus (Mircescu et al., 2009).

In the context of studying fungal infections, image-based systems biology is able to serve as a well-founded framework with iterative cycles of exchange between experiment and theory and involves imaging, quantitative characterization and modeling of infection processes (Medyukhina et al., 2015). Methods for image-analysis of fungal-host interactions (Mech et al., 2011; Kraibooj et al., 2014; Brandes et al., 2015) and parameter-free classification of cell-tracks (Mokhtari et al., 2013) have been developed over the recent years and have paved the way for the quantification and extraction of the information contained in image- and video data. Furthermore, different individual-based modeling approaches were successfully carried out in combination with automated image-analysis to test hypotheses and to draw predictions that might be tested in future experimental research (Tokarski et al., 2012; Mech et al., 2013; Hünniger et al., 2014). Experimental studies including live-cell imaging in alveolar ducts would give the opportunity to refine, to review and to extend the present virtual infection model.

Author Contributions

Conception and design of the investigation and work: JP, MTF. Contribution of materials and computational resources: MTF. Data processing, implementation and application of the computational algorithm: JP. Evaluation and analysis of the results: JP, MTF. Drafting the manuscript and revising it critically for important intellectual content and final approval of the version to be published: JP, MTF. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: JP, MTF.

Acknowledgments

This work was financially supported by the excellence graduate school Jena School for Microbial Communication (JSMC) and the CRC/TR124 FungiNet, Project B4, that are both funded by the Deutsche Forschungsgemeinschaft (DFG).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00503/abstract

References

Alonso, C., Waring, A., and Zasadzinski, J. A. (2005). Keeping lung surfactant where it belongs: protein regulation of two-dimensional viscosity. *Biophys. J.* 89, 266–273. doi: 10.1529/biophysj.104.052092

Bastacky, J., and Lee, C. (1995). Alveolar lining layer is thin and continuous: low-temperature scanning electron microscopy of rat lung. *J. Appl. Physiol.* 79, 1615–1628.

Behnsen, J., and Hartmann, A. (2008). The opportunistic human pathogenic fungus *Aspergillus fumigatus* evades the host complement system. *Infect. Immun.* 76, 820–827. doi: 10.1128/IAI.01037-07

Beyer, T., and Meyer-Hermann, M. (2008). Cell transmembrane receptors determine tissue pattern stability. *Phys. Rev. Lett.* 101:148102. doi: 10.1103/PhysRevLett.101.148102

Brakhage, A. A., Bruns, S., Thywissen, A., Zipfel, P. F., and Behnsen, J. (2010). Interaction of phagocytes with filamentous fungi. *Curr. Opin. Microbiol.* 13, 409–415. doi: 10.1016/j.mib.2010.04.009

Brandes, S., Mokhtari, Z., Essig, F., Hünniger, K., Kurzai, O., and Figge, M. T. (2015). Automated segmentation and tracking of non-rigid objects in time-lapse microscopy videos of polymorphonuclear neutrophils. *Med. Image Anal.* 20, 34–51. doi: 10.1016/j.media.2014.10.002

Brown, K. (1979). Voronoi diagrams from convex hulls. *Inf. Process. Lett.* 9:1979. doi: 10.1016/0020-0190(79)90074-7

De Berg, M., Cheong, O., Van Kreveld, M., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*, Vol. 17. (Berlin; Heidelberg: Springer-Verlag).

Devreotes, P. N., and Zigmond, S. H. (1988). Chemotaxis in eukaryotic cells: a focus on leukocytes and Dictyostelium. *Annu. Rev. Cell Biol.* 4, 649–686. doi: 10.1146/annurev.cb.04.110188.003245

Farrell, B. E., Daniele, R. P., and Lauffenburger, D. A. (1990). Quantitative relationships between single-cell and cell-population model parameters for chemosensory migration responses of alveolar macrophages to C5a. *Cell Motil. Cytoskeleton* 16, 279–293. doi: 10.1002/cm.970160407

Fels, A., and Cohn, Z. (1986). The alveolar macrophage. *J. Appl. Physiol.* 60, 353–369.

Francis, K., and Palsson, B. O. (1997). Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proc. Natl. Acad. Sci. U.S.A.* 94, 12258–12262. doi: 10.1073/pnas.94.23.12258

Guo, Z., Sloot, P. M. A., and Tay, J. C. (2008). A hybrid agent-based approach for modeling microbiological systems. *J. Theor. Biol.* 255, 163–175. doi: 10.1016/j.jtbi.2008.08.008

Guo, Z., and Tay, J. (2008). "Granularity and the validation of agent-based models," in *Proceedings of the 2008 Spring Simulation Multiconference* (San Diego, CA), 153–161.

Hasenberg, M., Behnsen, J., Krappmann, S., Brakhage, A., and Gunzer, M. (2011). Phagocyte responses towards *Aspergillus fumigatus*. *Int. J. Med. Microbiol.* 301, 436–444. doi: 10.1016/j.ijmm.2011.04.012

Hasenberg, M., Stegemann-Koniszewski, S., and Gunzer, M. (2013). Cellular immune reactions in the lung. *Immunol. Rev.* 251, 189–214. doi: 10.1111/imr.12020

Heinekamp, T., Schmidt, H., Lapp, K., Pähtz, V., Shopova, I., Köster-Eiserfunke, N., et al. (2014). Interference of *Aspergillus fumigatus* with the immune

response. *Semin. Immunopathol.* 37, 141–152. doi: 10.1007/s00281-014-0465-1

Herzog, E. L., Brody, A. R., Colby, T. V., Mason, R., and Williams, M. C. (2008). Knowns and unknowns of the alveolus. *Proc. Am. Thoracic Soc.* 5, 778–782. doi: 10.1513/pats.200803-028HR

Hohl, T. M. (2008). Stage-specific innate immune recognition of *Aspergillus fumigatus* and modulation by echinocandin drugs. *Med. Mycol.* 47(Suppl. I), 1–7. doi: 10.1080/13693780802078131

Horn, F., Heinekamp, T., Kniemeyer, O., Pollmächer, J., Valiante, V., and Brakhage, A. A. (2012). Systems biology of fungal infection. *Front. Microbiol.* 3:108. doi: 10.3389/fmicb.2012.00108

Hünniger, K., Lehnert, T., Bieber, K., Martin, R., Figge, M. T., and Kurzai, O. (2014). A virtual infection model quantifies innate effector mechanisms and candida albicans immune escape in human blood. *PLOS Comput. Biol.* 10:e1003479. doi: 10.1371/journal.pcbi.1003479

Iglesias, P. A., and Devreotes, P. N. (2008). Navigating through models of chemotaxis. *Curr. Opin. Cell Biol.* 20, 35–40. doi: 10.1016/j.ceb.2007.11.011

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Kozel, T. R., Wilson, M. A., Farrell, T. P., and Levitz, S. M. (1989). Activation of C3 and binding to *Aspergillus fumigatus* conidia and hyphae. *Infect. Immun.* 57, 3412–3417.

Kraibooj, K., Park, H.-R., Dahse, H.-M., Skerka, C., Voigt, K., and Figge, M. T. (2014). Virulent strain of Lichtheimia corymbifera shows increased phagocytosis by macrophages as revealed by automated microscopy image analysis. *Mycoses* 57, 56–66. doi: 10.1111/myc.12237

Krombach, F., and Münzing, S. (1997). Cell size of alveolar macrophages: an interspecies comparison. *Environ. Health Perspect.* 105, 1261–1263. doi: 10.1289/ehp.97105s51261

Lin, S., Schranz, J., and Teutsch, S. (2001). Aspergillosis case-fatality rate: systematic review of the literature. *Clin. Infect. Dis.* 32, 358–366. doi: 10.1086/318483

MacWilliam, T., and Cecka, C. (2013). "CrowdCL: web-based volunteer computing with WebCL," in *2013 IEEE High Performance Extreme Computing Conference (HPEC)* (Waltham, MA), 1–6.

Mech, F., Thywissen, A., Guthke, R., Brakhage, A. A., and Figge, M. T. (2011). Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS ONE* 6:e19591. doi: 10.1371/journal.pone.0019591

Mech, F., Wilson, D., Lehnert, T., Hube, B., and Thilo Figge, M. (2013). Epithelial invasion outcompetes hypha development during Candida albicans infection as revealed by an image-based systems biology approach. *Cytometry A.* 85, 126–139. doi: 10.1002/cyto.a.22418

Medyukhina, A., Timme, S., Mokhtari, Z., and Figge, M. T. (2015). Image-based systems biology of infection. *Cytometry A.* doi: 10.1002/cyto.a.22638. [Epub ahead of print].

Miller, A. L., Strieter, R. M., Gruber, A. D., Ho, S. B., and Lukacs, N. W. (2003). CXCR2 regulates respiratory syncytial virus-induced airway hyperreactivity and mucus overproduction. *J. Immunol.* 170, 3348–3356. doi: 10.4049/jimmunol.170.6.3348

Mircescu, M. M., Lipuma, L., van Rooijen, N., Pamer, E. G., and Hohl, T. M. (2009). Essential role for neutrophils but not alveolar macrophages at early time points following *Aspergillus fumigatus* infection. *J. Infect. Dis.* 200, 647–656. doi: 10.1086/600380

Mokhtari, Z., Mech, F., Zitzmann, C., Hasenberg, M., Gunzer, M., and Figge, M. T. (2013). Automated characterization and parameter-free classification of cell tracks based on local migration behavior. *PLoS ONE* 8:e80808. doi: 10.1371/journal.pone.0080808

O'Gorman, C. M., and Fuller, H. T. (2008). Prevalence of culturable airborne spores of selected allergenic and pathogenic fungi in outdoor air. *Atmos. Environ.* 42, 4355–4368. doi: 10.1016/j.atmosenv.2008.01.009

Pelletier, A. (2000). Presentation of chemokine SDF-1α by fibronectin mediates directed migration of T cells. *Blood* 96, 2682–2690.

Pollmächer, J., and Figge, M. T. (2014). Agent-based model of human alveoli predicts chemotactic signaling by epithelial cells during early *Aspergillus fumigatus* infection. *PLoS ONE* 9:e111630. doi: 10.1371/journal.pone.0111630

Randolph, G. J., Angeli, V., and Swartz, M. A. (2005). Dendritic-cell trafficking to lymph nodes through lymphatic vessels. *Nat. Rev. Immunol.* 5, 617–628. doi: 10.1038/nri1670

Riedemann, N. C., Guo, R.-F., Sarma, V. J., Laudes, I. J., Huber-Lang, M., Warner, R. L., et al. (2002). Expression and function of the C5a receptor in rat alveolar epithelial cells. *J. Immunol.* 168, 1919–1925. doi: 10.4049/jimmunol.168.4.1919

Sbalzarini, I. F., Hayer, A., Helenius, A., and Koumoutsakos, P. (2006). Simulations of (an)isotropic diffusion on curved biological surfaces. *Biophys. J.* 90, 878–885. doi: 10.1529/biophysj.105.073809

Sklar, L. A. (1984). The dynamics of ligand-receptor interactions. *J. Biol. Chem.* 259, 5661–5669.

Sukumar, N. (2003). Voronoi cell finite difference method for the diffusion operator on arbitrary unstructured grids. *Int. J. Numer. Methods Eng.* 57, 1–34. doi: 10.1002/nme.664

Thomson, J. (1904). XXIV. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philos. Mag. Ser.* 7, 237–265. doi: 10.1080/14786440409463107

Tokarski, C., Hummert, S., Mech, F., Figge, M. T., Germerodt, S., Schroeter, A., et al. (2012). Agent-based modeling approach of immune defense against spores of opportunistic human pathogenic fungi. *Front. Microbiol.* 3:129. doi: 10.3389/fmicb.2012.00129

Tranquillo, R., Fisher, E., Farrell, B., and Lauffenburger, D. (1988). A stochastic model for chemosensory cell movement: application to neutrophil and macrophage persistence and orientation. *Math. Biosci.* 90, 287–303. doi: 10.1016/0025-5564(88)90071-5

van Haastert, P. J. M., and Postma, M. (2007). Biased random walk by stochastic fluctuations of chemoattractant-receptor interactions at the lower limit of detection. *Biophys. J.* 93, 1787–1796. doi: 10.1529/biophysj.107.104356

van Lookeren Campagne, M., Wiesmann, C., and Brown, E. J. (2007). Macrophage complement receptors and pathogen clearance. *Cell. Microbiol.* 9, 2095–2102. doi: 10.1111/j.1462-5822.2007.00981.x

Zipfel, P. F., and Skerka, C. (2009). Complement regulators and inhibitory proteins. *Nat. Rev. Immunol.* 9, 729–740. doi: 10.1038/nri2620

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 09 June 2015
doi: 10.3389/fmicb.2015.00549

Automated quantification of the phagocytosis of *Aspergillus fumigatus* conidia by a novel image analysis algorithm

Kaswara Kraibooj[1,2†], *Hanno Schoeler*[2,3†], *Carl-Magnus Svensson*[1], *Axel A. Brakhage*[2,3] *and Marc Thilo Figge*[1,2*]

[1] *Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany,* [2] *Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, Jena, Germany,* [3] *Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Jena, Germany*

OPEN ACCESS

Edited by:
Tunahan Cakir, Gebze Technical University, Turkey

Reviewed by:
Frederic Lamoth, Lausanne University Hospital, Switzerland
Alfonso Caiazzo, Weierstrass Institute for Applied Analysis and Stochastics, Germany

***Correspondence:**
Marc Thilo Figge, Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology – Hans Knöll Institute, Beutenbergstrasse 11a, 07745 Jena, Germany
thilo.figge@hki-jena.de

† *These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *15 April 2015*
Accepted: *19 May 2015*
Published: *09 June 2015*

Citation:
Kraibooj K, Schoeler H, Svensson C-M, Brakhage AA and Figge MT (2015) Automated quantification of the phagocytosis of Aspergillus fumigatus conidia by a novel image analysis algorithm. Front. Microbiol. 6:549. doi: 10.3389/fmicb.2015.00549

Studying the pathobiology of the fungus *Aspergillus fumigatus* has gained a lot of attention in recent years. This is due to the fact that this fungus is a human pathogen that can cause severe diseases, like invasive pulmonary aspergillosis in immunocompromised patients. Because alveolar macrophages belong to the first line of defense against the fungus, here, we conduct an image-based study on the host-pathogen interaction between murine alveolar macrophages and *A. fumigatus*. This is achieved by an automated image analysis approach that uses a combination of thresholding, watershed segmentation and feature-based object classification. In contrast to previous approaches, our algorithm allows for the segmentation of individual macrophages in the images and this enables us to compute the distribution of phagocytosed and macrophage-adherent conidia over all macrophages. The novel automated image-based analysis provides access to all cell-cell interactions in the assay and thereby represents a framework that enables comprehensive computation of diverse characteristic parameters and comparative investigation for different strains. We here apply automated image analysis to confocal laser scanning microscopy images of the two wild-type strains ATCC 46645 and CEA10 of *A. fumigatus* and investigate the ability of macrophages to phagocytose the respective conidia. It is found that the CEA10 strain triggers a stronger response of the macrophages as revealed by a higher phagocytosis ratio and a larger portion of the macrophages being active in the phagocytosis process.

Keywords: *Aspergillus fumigatus*, alveolar macrophages, host-pathogen interaction, phagocytosis assay, automated image analysis

Introduction

The use of sophisticated microscopy techniques and the ease of producing and storing large amount of image data has in the last decade led to an increasing need for automated image analysis tools that reveal and quantify biological processes on a systems biology level (Rittscher, 2010). We will here present an automated image analysis algorithm that is able to fully evaluate the data from a phagocytosis assay between murine alveolar macrophages and the fungus *Aspergillus fumigatus*.

This is a biologically relevant experiment as the ubiquitous saprophytic mold *A. fumigatus* is the most prevalent airborne fungal pathogen (Brakhage et al., 1999; Brakhage and Langfelder, 2002; O'Gorman et al., 2008). During its asexual reproduction cycle the fungus produces conidia that are inhaled by humans at a rate of hundreds to thousands per day, without any consequences for healthy humans (Latgé, 1999). In immunocompromised subjects, by contrast, *A. fumigatus* can cause invasive pulmonary aspergillosis (IPA) which has mortality rates in the order of 30–95% (Brakhage, 2005; Dagenais and Keller, 2009). From the host side, macrophage phagocytosis is part of the early response of the innate immune system and the igniting process of the adaptive immune system at a later stage (Aderem and Underhill, 1999). Therefore, *in vitro* phagocytosis assays where *A. fumigatus* conidia are confronted with mammalian macrophage cells are a suitable experiment to examine the interaction between pathogen and host, thereby gaining deeper insight into mechanisms of pathogenicity and phagocytosis.

The evaluation of phagocytosis assays based on images is often carried out by visual inspection and is therefore time consuming, subjective and expensive, accentuating the need for a more efficient analysis method (Zhou and Wong, 2006). The automated image analysis performed in this work was realized within the *Definiens Developer XD* framework (Schönmeyer et al., 2011) that allows creating customized algorithms tailored for the life sciences (Carpenter et al., 2006) and facilitates automated analysis of big sets of image data by batch processing. Additionally, this platform enables to conveniently combine predefined features of image objects by mathematical and logical combinations. Studying biological phenomena often requires analyzing microscopy images with a high degree of variation in object features, both across the images and even across objects in the same image. A clear example in the present context is the variation in shape of clusters formed by different numbers of conidia whose relative positions in the clusters are different for virtually all clusters. Another example is the change of the mean intensity of objects depending on how deep a cell lies in the experimental well. Validation of our algorithm ensures its applicability to a wide spectrum of image data required for high-throughput screening of mutants in comparative studies.

Here, we performed confocal laser scanning microscopy (CLSM) experiments to study two clinical isolates of *A. fumigatus*: ATCC 46645 (American Type Culture Collection, Manassas, VA) and CEA10. These two strains, among others, were examined for virulence in an embryonated egg model and CEA10 displayed an increased virulence compared with ATCC (Jacobsen et al., 2010). It should be noted that in this study older embryos had an increased survival chance and this was hypothesized to be caused by a more mature immune system. By performing the confrontation assay carried out here, we wanted to elucidate the infection process of the two strains and shed light on the mechanisms of CEA10's virulence, especially its interaction with a mature immune system. In our experiment we compared the phagocytosis ratio, macrophage-adherence and aggregation of the two strains.

In the microscopy experiments performed for the present study, different fluorescent dyes for macrophages and conidia were used and, in addition, the technique of differential staining was applied to distinguish phagocytosed from non-phagocytosed conidia (Thywißen et al., 2011). In the standard red-blue-green (RGB) formulation of a color image, each color layer displays a specific class of cells, i.e., all macrophages in the red layer, all conidia in the green layer and all non-phagocytosed conidia in the blue layer. Hence, the staining protocol enabled us to conveniently work on single layers for object segmentation and to ultimately combine layers in the classification of objects and in the analysis of their spatial colocalizations. While some progress has been made in previous developments of algorithms for the automated image analysis of this type of experiments (Mech et al., 2011, 2014; Kraibooj et al., 2014; Schäfer et al., 2014), we are here presenting a novel algorithm that differs from previous approaches with regard to the crucial step of segmenting all different types of cells in the assay. This was not achieved in previous approaches regarding the segmentation of macrophages, which is complicated by the occasionally interrupted staining of their cell surface. Since reliable segmentation of all cells in the assay is a necessary prerequisite for its comprehensive quantification, we could go beyond previous work and computed various biological quantities, such as the phagocytic index (Sano et al., 2003) and other measures involving the quantification of macrophages, which was not possible in previous studies.

Materials and Methods

A. fumigatus Strains and Growth Condition

Cultivation of the *A. fumigatus* wild-type ATCC 46645 and CEA10 was performed on *Aspergillus* minimal medium (AMM) agar plates with 1% (w/v) glucose at 37°C for 5 days. AMM consisted of 70 mM $NaNO_3$, 11.2 mM KH_2PO_4, 7 mM KCl, 2 mM $MgSO_4$ and 1 µl/ml trace element solution (pH 6.5). The trace element solution consisted of 1 g $FeSO_4 * 7\ H_2O$, 8.8 g $ZnSO_4 * 7\ H_2O$, 0.4 g $CuSO_4 * 5\ H_2O$, 0.15 g $MnSO_4 * H_2O$, 0.1 g $NaB_4O_7 * 10\ H_2O$, 0.05 g $(NH_4)_6Mo_7O_{24} * 4\ H_2O$, and double-distilled water (ddH_2O) to 1000 ml (Brakhage and Van den Brulle, 1995; Maerker et al., 2005). Conidia were harvested in sterile, double-distilled water.

Phagocytosis Assays and Cell Staining

For the phagocytosis assays murine alveolar macrophages (ATCC CRL-2019™) were cultivated in RPMI1640 medium supplemented with 10% (v/v) FCS (Thermo Fisher Scientific, Dreieich, Germany), 1% (w/v) sodium bicarbonate (Lonza, Köln, Germany) and 0.05 mM beta-mercaptoethanol (Life Technologies, Darmstadt, Germany). The cells were seeded on glass cover slips in Nunc 24 well plates (Thermo Scientific) at a density of 3×10^5 cells per well and allowed to grow adherently overnight. The conidia were stained with Fluorescein isothiocyanate (FITC, Sigma, Taufkirchen, Germany) for 30 min at 37°C while shaking. After washing them 3 times with PBS, 0.01% (v/v) Tween20 (AppliChem, Darmstadt, Germany) conidia concentration was determined using a CASY cell counter model TT (Roche-Innovatis, Penzberg, Germany). Conidia were added to the macrophages at a multiplicity of infection

(MOI) of 7. Synchronization of the experiment was realized by centrifugation for 5 min at 100 g and 37°C. To initiate the experiment the co-incubation was shifted to a humidified CO_2 incubator for 1 h at 37°C. The cells were fixed for 15 min at room temperature by adding 16% (v/v) paraformaldehyde (Electron Microscopy Sciences, München, Germany) directly to the medium to a concentration of 3.7% (v/v). Wells were washed two times with PBS followed by the step of differential staining, i.e., non-phagocytosed conidia were stained with 0.1 mg/ml calcofluor white (Sigma) for 30 min at room temperature. The cells were washed again twice with PBS. Prior to antibody labeling, binding sites were blocked with PBS, 3% (w/v) BSA Fraktion V (AppliChem) for 30 min. Next, macrophages were labeled with a monoclonal rat anti-CD9 antibody (1:200; Santa Cruz Biotechnology, Heidelberg, Germany) over night at 4°C and an Alexa Fluor® 647 Goat Anti-Rat IgG antibody (1:200; Life Technologies) for 1.5 h at room temperature.

Imaging

Microscopy images were taken by a Zeiss LSM 780 Live confocal laser scanning microscope with a 20x Zeiss plan-apochromat dry objective (0.8 NA). The resulting images are 8-bit RGB color images with 1024 × 1024 pixels and a pixel distance of 0.2 µm. The total number of images for each strain is 60, equally divided into two biological replicates (i.e., 30 images each) with two technical replicates per biological replicate (i.e., 15 images each). These image data are publicly available at http://www.leibniz-hki.de/en/asb-downloads.html. Based on the differential staining, all macrophages appeared in the red layer, all conidia in the green layer and all non-phagocytosed conidia were visible in the blue layer (see **Figure 1**). The separation of objects into different layers allowed for more effective segmentation and classification of the different image objects.

Automated Image Analysis

The algorithm for automated image analysis was developed using the software *Definiens Developer XD* and was executed by the *Grid XD Server* (Definiens AG, Munich, Germany). The server ran on one core of a SUN Fire X4600 Server M2 (8 CPUs with 4 cores each, 2.3 GHz AMD Opteron, 64 GB memory). Image processing consisted of three subsequent steps—preprocessing, segmentation and classification—and a schematic overview of the algorithm is presented in **Figure 2**. The rule set of the algorithm is provided as Supplementary Material (see Supplementary Data 1) and the code is available by the authors upon request.

Preprocessing

As shown in **Figure 2** (see split point 1) the green and red layer were separated and smoothed by a Gaussian filter, i.e., a low-pass filter that reduces high-frequency components in the image (Blinchikoff and Krause, 2001). The degree of smoothing by the Gaussian filter was adjusted through the parameter σ, which controls the standard deviation. This was chosen as to optimize the subsequent image segmentation: in the green layer $\sigma = 1$ px was applied and in the red layer $\sigma = 5$ px was used. For the green layer a relatively small σ was sufficient as the conidia were intensity-wise homogeneous and well separated from the background (see **Figure 1A**). In contrast, the higher value of σ applied on the red layer alleviated conspicuous discontinuities stemming from the staining of macrophages. Blurring helped bridging these discontinuities in macrophage boundaries and consequently improved the subsequent segmentation. Preprocessing of the blue layer, which exclusively indicated non-phagocytosed conidia, was not required, because segmentation was not applied to this layer. The calcofluor white signal, represented by the pixel intensities in this layer, was used for classification of conidia as phagocytosed or not. **Figures 3A,B**, **4A,B** show the smoothing effects on the green and red layers, respectively; this effect was very prominent for macrophages in the red layer.

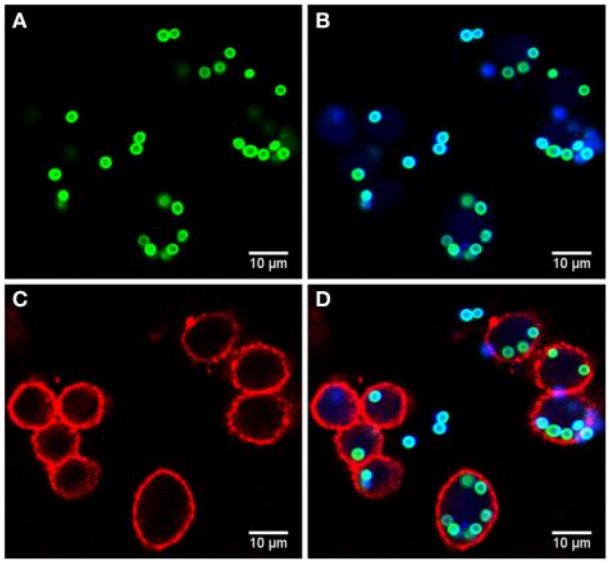


FIGURE 1 | Example of image data from the phagocytosis assay. (A) Green layer (FITC staining) with all conidia. **(B)** Blue layer (Calcofluor white staining) overlaid with the FITC layer revealing the difference between phagocytosed and non-phagocytoced conidia as only the latter were stained with Calcofluor white (differential staining). **(C)** Red layer (anti-CD9 antibody) with macrophages and **(D)** overlay of all layers.

Segmentation

Segmentation of conidia

Segmenting conidia on the green layer comprised two consecutive phases. In the first phase, indicated by split point 2 in **Figure 2**, the image was segmented into background and foreground (conidia candidates) using clustering-based thresholding (Sezgin and Sankur, 2004). The threshold, T_c, was automatically computed by a combination of intensity histogram-based measures and homogeneity measurements of segmented objects (Pal and Pal, 1993). An example of this thresholding procedure is shown in **Figures 3B,C**. Note that the resulting foreground objects are either single conidia, cluster of conidia or debris, which were further distinguished in the second phase of conidia segmentation. After thresholding, a morphological closing operation was performed (Gonzalez

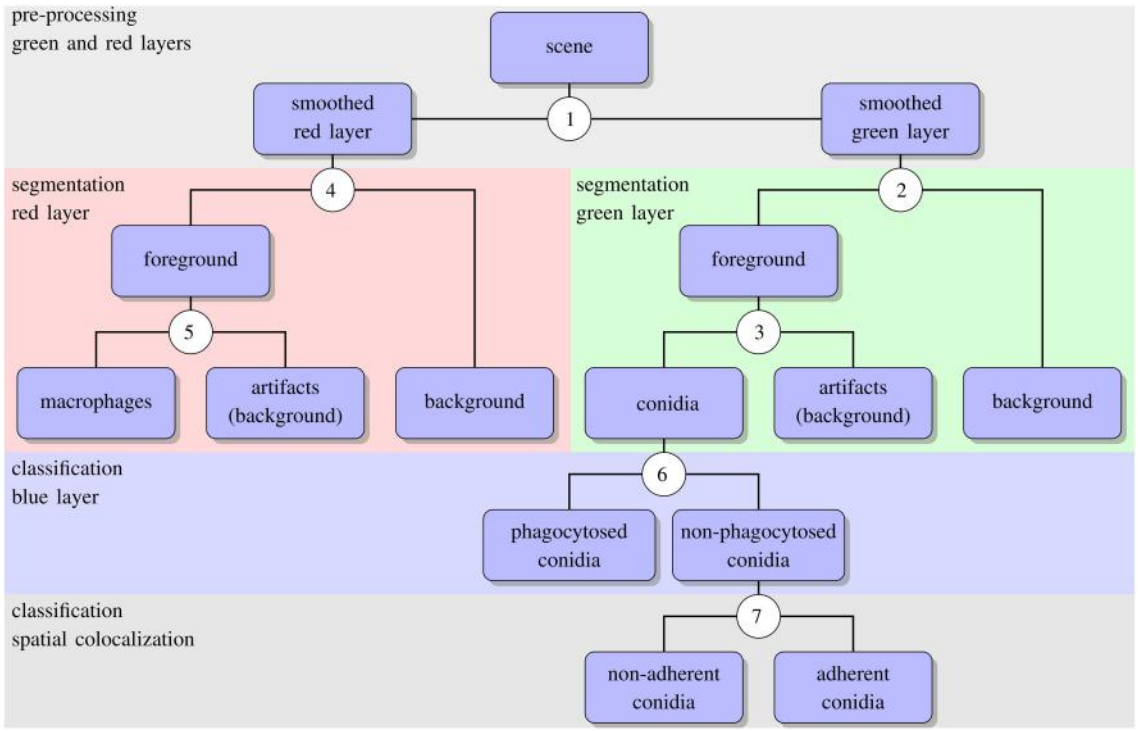


FIGURE 2 | Schematic representation of the automated image analysis algorithm with boxes showing input and output of image analysis processes. Circled numbers indicate split points where the different processes were performed, e.g., thresholding or object classification. For the details of each split point we refer the reader to the text.

and Woods, 2006) to avoid having background enclosed in conidia as this is not a biologically realistic scenario. At the heart of identifying individual foreground objects is the *watershed segmentation*, which is a standard method for segmenting objects based on seed points that lie inside the desired objects. Starting from these seed points, it can be imagined that the intensity landscape is flooded with water filling up each object until it meets water from another object. At locations where water from different basins meets, a dam is placed that separates individual objects from each other and by that yields their segmentation (Roerdink and Meijster, 2001). In our case, seed points for each foreground object were obtained by first applying a distance transform (Fabbri et al., 2008) to the local maxima in the intensity landscape.

In the second phase, a foreground object was considered to be a cluster if its area, $A_{cl} \geq A_{cl}^{min} = 275$ px, otherwise it was considered to be a single conidium. To split a cluster into single conidia, we iteratively applied the distance transform (Fabbri et al., 2008) on the cluster to obtain seed points that were then used in the subsequent watershed (see **Figure 3E**). The seed points were chosen to be the local maxima obtained from the distance transform. Each time before watershed segmentation was carried out, the considered cluster was shrunk to facilitate segmentation of individual conidia. The shrinking procedure was realized by discarding pixels from the old border that have intensity below a given threshold. Before watershed segmentation was applied, all clusters were shrunk with threshold $T_{shrink}^{global} = 40$ and after that, individual clusters were handled by initializing T_{shrink}^{cl} as the minimal intensity of a specific cluster under consideration. The individual threshold was iteratively updated by $T_{shrink}^{cl} \longleftarrow T_{shrink}^{cl} + I_{inc}$, where $I_{inc} = 10$ was set during the engineering process. The iteration for splitting clusters into single conidia stopped when all segments in the original cluster had area $A_{cl} < A_{cl}^{min}$. The shrinking procedure was required to enhance the performance of watershed segmentation with regard to (i) identifying single conidia in clusters and (ii) correcting for a single conidium that was previously erroneously identified as a cluster. In the first case, the shrinking supported the determination of the borders of single conidia, which was necessary to identify suitable seeds for watershed segmentation. The shrinking excluded halos and reflections, which were especially strong for large clusters, and yielded seeds located on the actual conidia. Without shrinking, some seeds would have been placed with high likelihood on halos and reflections, which would lead to over-segmentation of a cluster. In the second case, the object was reduced in area below the threshold A_{cl}^{min}. This procedure corrected for single conidia with areas above A_{cl}^{min} that can occur due to a halo caused by reflections or due to another out-of-focus conidium in its close proximity. The iteration automatically stopped when no objects were classified as clusters.

Image objects with area $A_c \leq A_c^{min} = 100$ px, or roundness $\rho \geq \rho_c^{max} = 1.2$ were discarded, because the engineering process revealed that these were most likely artifacts. Here, roundness

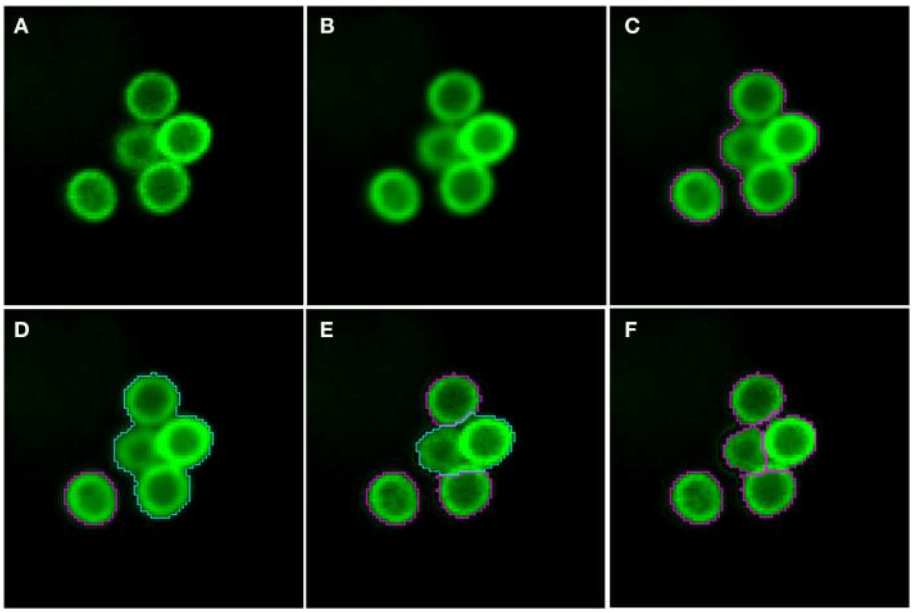


FIGURE 3 | Pre-processing and segmentation of conidia. (A) Original scene showing a cluster of conidia and an isolated conidium. **(B)** Scene after pre-processing with a Gaussian filter. **(C)** Scene after thresholding into background and foreground indicated by magenta borders. **(D)** Distinguishing the foreground between clusters (cyan) and single conidia (magenta). Segmentation of the cluster into smaller clusters and finally single conidia after **(E)** the first and **(F)** the third iteration.

was measured as $\rho = \varepsilon_v^{max} - \varepsilon_v^{min}$, where ε_v^{max} is the major radius of the smallest enclosing ellipse of image object v and ε_v^{min} is the minor radius of the largest ellipse enclosed by v. Accordingly, $\rho \in [0, \infty)$ and for $\rho = 0$ the object has a perfectly circular shape. Additionally, image objects of mean green intensity $I_c \leq I_c^{min} = 20$ were also discarded, because they were in fact out of focus. Moreover, image objects were discarded if the ratio of the long over the short main axis was bigger than 2. The iterative segmentation and exclusion of debris were the processes of split point 3 in **Figure 2**. An example for the segmentation of conidia in a cluster is shown in **Figures 3D–F**.

Segmentation of macrophages

The antibody-stained macrophages in the red layer were segmented in a way similar to the segmentation of conidia although three distinct phases had to be distinguished in this case: (i) thresholding, (ii) watershed segmentation and (iii) morphology-based macrophage identification. Firstly, the preprocessed layer of macrophages was segmented into background and foreground (split point 4 in **Figure 2**) using an automatic threshold applied to the preprocessed layer. This threshold was multiplied by a factor $f = 0.3$ in order not to lose any macrophage signal, as shown in **Figures 4B,C**. The resulting background with intensity below threshold consisted of regions in-between and outside macrophages, as well as areas inside the macrophages. The thresholded foreground consisted of macrophages and possibly adjacent background with pixels of high intensities due to halo effects and smoothing. Next, background segments enclosed by foreground segments were merged into the foreground. This ensured that the application of watershed segmentation performed well, because the intensity minima in these enclosed segments were used as seeds for watershed segmentation (see **Figures 4C,D**).

Secondly, the foreground had to be divided into regions corresponding to individual macrophages or unwanted background. To achieve this we applied an edge detection filter and then smoothed the foreground with a Gaussian filter of width $\sigma = 9$ px and applied watershed segmentation (see **Figures 4E,F**). Finally we identified the segments given by watershed segmentation as either macrophages or background. We could distinguish between two types of background segments according to their positions: macrophage-adjacent background segments that are bordering the background determined in the thresholding phase of segmentation and background segments that were completely enclosed by macrophages and had no contact with the background of the first phase. All macrophage-adjacent background segments were merged with the background from the thresholding, see **Figure 4F**. However, the enclosed background segments could not be classified in this way, because they had no contact with the background of the thresholding and determining whether these were macrophages or background required morphology-based macrophage identification using a rule-based classifier. The macrophage identification rules were based on the object's shape and brightness. The brightness β is the average intensity over the object's pixels and we consider the shape properties roundness ρ and shape index ς. The shape index describes the smoothness of an object border and is defined as $\varsigma = b_v/(4\sqrt{P_v})$, where b_v is the border length of image object v and P_v is the number of all

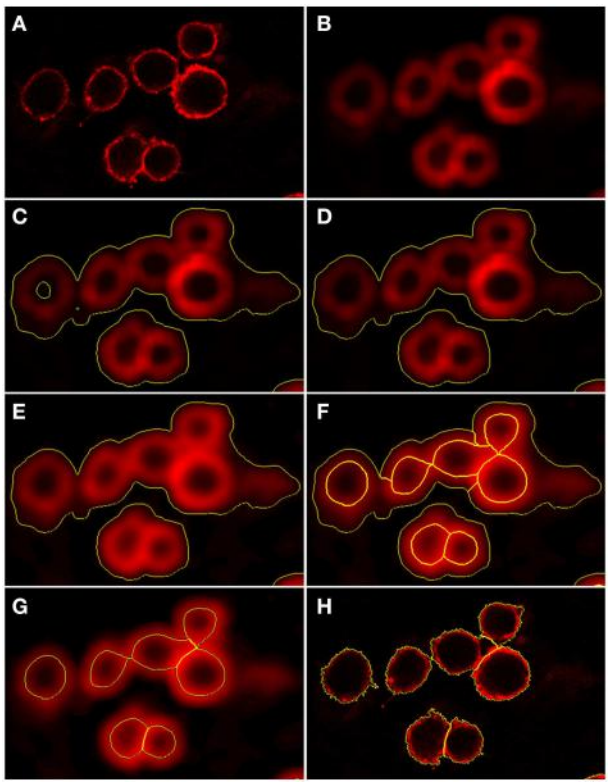


FIGURE 4 | Preprocessing and segmentation of macrophages. (A) Original scene showing macrophages. **(B)** Scene after pre-processing with a Gaussian filter. **(C)** Scene after thresholding into background and foreground indicated by yellow borders. **(D)** Scene after merging the background enclosed by foreground into foreground. **(E)** Scene after edge detection and application of a Gaussian filter. Scene after **(F)** application of watershed segmentation and **(G)** merging of surrounding border foreground into background. **(H)** Scene after region growing enabling classification into individual macrophages.

pixels forming this object. This implies $\varsigma \in [1, \infty)$ and $\varsigma = 1$ indicates perfectly smooth borders. Together with the roundness ρ, these three macrophage features are considered separately as well as in various combinations to achieve optimal segmentation results. Note that we are considering two separate populations of segments when applying this ruleset, namely segments that are in contact with the image border and those which are not. For segments in contact with the border the measures have different thresholds when deciding whether they are macrophages or background.

The borders of the resulting macrophage segments are mostly located at the inner borders of macrophages in the image (see **Figure 4G**). To obtain the outer border we apply a region growing algorithm that first is expanding the border a fixed number of pixels in the radial direction. The engineering process revealed that two pixels was a suitable value for this growing process. Subsequently the region was grown pixel-wise in the radial direction until a step was made with an intensity drop by more than 30%. The growing process was restricted by a roundness condition preventing the macrophage segment to be extended into neighboring macrophages. This condition was expressed by the roundness $\rho_m^* \leq \rho_m + 0.01$, where ρ_m is the initial roundness and ρ_m^* is the roundness after growing. Finally, macrophage segments with area A_m below the threshold value $A_m^{min} = 2400$ px were discarded because these did most likely represent artifacts (see split point 5 in **Figure 2**). Examples of final segmentation results are shown in **Figure 4H**.

Classification

The segmentation of image objects was followed by their classification. In particular, for each conidium we had to determine whether or not it was phagocytosed, and if not phagocytosed whether or not it was adherent to a macrophage. To distinguish between phagocytosed and non-phagocytosed conidia, we exploited the information provided from the differential staining of conidia. For each conidium, the number of calcofluor white signal was measured by computing the average blue intensity, I_c^{blue}, where a lower (higher) value than the threshold intensity T_c^{blue} indicated that the conidium was phagocytosed (non-phagocytosed) (split point 6 in **Figure 2**). By validation of the automated image analysis in comparison with a manual analysis, we inferred that $T_c^{blue} = 37$ is a suitable threshold value.

Next, for non-phagocytosed conidia, we differentiated between adherent and non-adherent conidia based on the relative position of conidia and macrophages (see split point 7 in **Figure 2**). To this end we computed the relative common border,

$$\Psi = \frac{\sum_{m \in N_c} b(c, m)}{b_c},$$

with $b(c, m)$ the length of the common border for objects c (conidium) and m (macrophages). N_c denotes the set of neighboring macrophages m relative to the conidium c and b_c denotes the border length of this conidium. It follows that $\Psi \in [0, 1]$, where $\Psi = 0$ implies that no common border exists between the conidum and a macrophage. In this case the conidium was not adherent to macrophages, whereas for $\Psi > 0$ the conidium had some common border with at least one macrophage and was therefore considered as being adherent. Moreover, we could use this equation to associate a specific conidium with a specific macrophage. A typical example of an image after performing classification is shown in **Figure 5**.

In close analogy to the above procedure, we could distinguish between an isolated conidium and conidia that occurred in aggregates. In this case, Ψ was computed among conidia and it was checked for common boundaries among them.

Statistical Analysis

We performed statistical tests to evaluate the statistical significance of our results using the Wilcoxon rank-sum test (Wilcoxon, 1945) and indicated the range of the p-value by n.s. for non-significant ($p \geq 0.05$), $*$ for $p < 0.05$, $**$ for $p < 0.01$, and $***$ for $p < 0.001$. Distributions of data points were represented by notched box plots (Krzywinski and Altman, 2014) representing the mean value (star), the median (horizontal line), boxes containing 25% of data points above and beyond the median, whiskers excluding 2% of data points above and beyond as possible outliers. Notches represent the

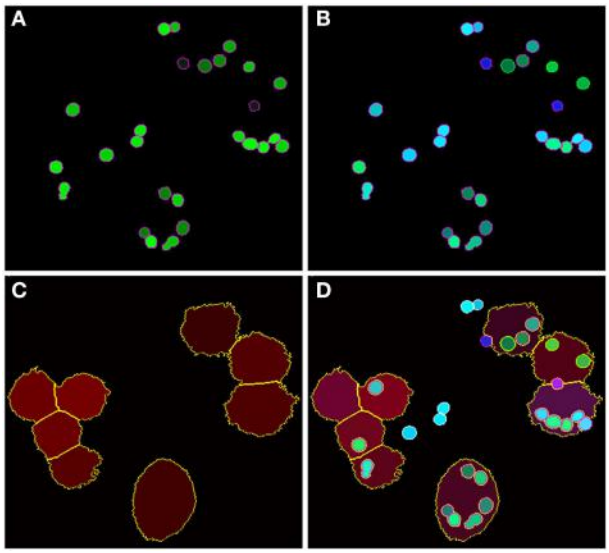


FIGURE 5 | Final image analysis result of the scene in Figure 1. (A) All conidia and **(B)** conidia after classification as being phagocytosed (green border) or non-phagocytosed (magenta border). **(C)** Segmented macrophages. **(D)** Overlay of all image layers showing the final result of image analysis.

95% confidence interval for their respective medians (Chambers, 1983).

Results

In this section, we first present results on the validation of the algorithm for the automated image analysis, which was done separately for macrophage segmentation and conidia identification, i.e., involving their segmentation and classification. Next, we provide the results from the quantification of the images in terms of the distribution of conidia over macrophages. Finally, the phagocytosis of conidia as well as their aggregation were quantitatively evaluated.

Automated Image Analysis Algorithm Reaches High Performance Measures

We validated our algorithm for the automated image analysis by comparison with an analysis that was carried out manually. To this end, a set of 24 images—i.e., 20% of the total number of images—were chosen randomly, such that each technical replicate was represented by three images. The visual inspection by experts was considered to be the ground truth, and the notions true positives (*TP*), false positives (*FP*) and false negatives (*FN*) were used accordingly to compute standard performance measures for binary classifications. The sensitivity is defined by

$$S = \frac{TP}{TP + FN},$$

whereas the precision is associated with the ratio

$$P = \frac{TP}{TP + FP}$$

and the accuracy is given by

$$A = \frac{TP}{TP + FP + FN},$$

where we set true negatives (*TN*) equal to zero, because these cannot occur in the current setting. All three performance measures can take values between 0 and 1, where high values indicate high performance with regard to the respective measure. The validation of macrophage segmentation as well as the identification of conidia revealed high performance measures and the results are summarized in **Table 1**.

In the case of macrophages, *TP* are the number of correctly segmented macrophages, *FN* are those that were erroneously considered to be background, whereas regions of background that were identified as macrophages are *FP*. *FP* arise from background regions in-between macrophages that happened to have shapes with roundness and shape index similar to actual macrophages. On the other hand, *FN* arise from macrophages that were not detected by the algorithm because of highly uneven staining and/or low integrity of the macrophage boundary, or because the roundness and shape index were similar to background segments, which can occur when a macrophage is partially covered by another macrophage.

The identification of conidia refers to the combined process of conidia segmentation and classification. In particular, we focused on the classification of phagocytosed vs. non-phagocytosed conidia. Thus, *TP* are the number of conidia which were correctly segmented and classified as either phagocytosed or non-phagocytosed. The conidia that were falsely identified—i.e., either erroneously segmented as conidia or falsely classified—were counted as *FP*, whereas *FN* are conidia that were missed at the level of segmentation.

In passing we note that we checked the technical and biological replicates of images for the MOI. In the experimental protocol, where the MOI was initially set to 7, several washing steps were required that affect the number of conidia relative to the number of macrophages. This loss of non-phagocytosed and mostly non-adherent conidia could not be avoided, however, we checked that the ultimate MOI was comparable in the images and found 2.42 ± 0.64 for the wild-type ATCC strain and 2.26 ± 0.76 for the CEA10 strain.

Distribution of Conidia Over Macrophages Yields Full Quantification of Image Data

We exploited the segmentation of macrophages to study the distribution of phagocytosed and adherent conidia over macrophages and the correlation between them. In **Figure 6** the probability distribution of macrophages as function of the number of adherent and the number of phagocytosed conidia is shown. This distribution contains the full information about the phagocytosis seen in the images. Already at this stage, qualitative differences between the strains when confronted with macrophages were detected. For example, comparing adherence of conidia to macrophages, it was observed this

occurred less frequently for the strain CEA10 than for ATCC. Approximately 25% of the macrophages had three or more adherent CEA10 conidia while more than 35% had three or more adherent ATCC conidia. Conversely, about 80% of the macrophages confronted with the ATCC strain did not phagocytose, whereas approximately 60% of the macrophages did not phagocytose when confronted with the CEA10 strain. This tendency yielded the qualitative information that immune cells responded more vigorously to CEA10 than to the ATCC strain. By summing the two-dimensional distribution along its individual axes we obtained the distribution of phagocytosed conidia and the distribution of adherent conidia (see **Figure 7**). These distributions confirmed the impression from **Figure 6** as the distribution of adherent conidia for ATCC was clearly shifted to higher conidia numbers compared to CAE10 (see **Figure 7A**), while the opposite was found in the case of phagocytosis (see **Figure 7B**).

Quantification of Phagocytosis Events Can Be Studied from Different Viewpoints

In previous studies, different scalar measures were used to quantify the phagocytosis process, which was partly a consequence of the fact that not all cells and interactions could be resolved (Sano et al., 2003; Mech et al., 2011; Kraibooj et al., 2014). We here present a comprehensive collection of various measures and compute them for comparison, i.e., from either the viewpoint of conidia or macrophages or from a combination of both viewpoints.

The conidia point of view is represented by the phagocytosis ratio that compares all phagocytosed conidia to all macrophage-associated conidia,

$$\varphi_c = \frac{N_c^{phag}}{N_c^{phag} + N_c^{adh}}.$$

Here, N_c^{phag} denotes the number of phagocytosed conidia and N_c^{adh} is the number of adherent conidia. It should be noted that this viewpoint intentionally neglects conidia that were not phagocytosed and not adherent to macrophages, because those conidia may have never been in contact with macrophages during the experiment. The phagocytosis ratio for the two strains is compared in **Figure 8A**, where the difference in phagocytosis, which was qualitatively discussed based on the distributions in **Figures 6**, **7**, was tested for statistical significance.

The macrophage point of view is expressed by the uptake ratio (Sano et al., 2003), which is the ratio of phagocytosing macrophages to all macrophages,

$$\varphi_m = \frac{N_m^{phag}}{N_m},$$

where N_m^{phag} denotes the number of phagocytosing macrophages and N_m the total number of macrophages. **Figure 8B** shows the uptake ratio for both strains. Although the phagocytosis ratio and the uptake ratio give a good idea about phagocytosis events in experiments, the picture is more complete when studying the mutual effect of both points of view. For this, the phagocytic index,

TABLE 1 | Validation of conidia identification and macrophage segmentation.

Process	Number	*TP*	*FP*	*FN*	S [%]	P [%]	A [%]
Macrophage segmentation	1130	1091	39	58	94.9	96.5	91.8
Conidia identification	2745	2697	48	76	97.2	98.2	95.6

Macrophage segmentation: true positives, TP, denote correctly segmented macrophages; false positives, FP, denote objects erroneously segmented as macrophages; false negatives, FN, denote macrophages missed in the segmentation. Conidia identification: true positives, TP, denote correctly segmented and classified conidia; false positives, FP, denote conidia that were falsely identified; false negatives, FN, denote conidia missed at the level of segmentation.

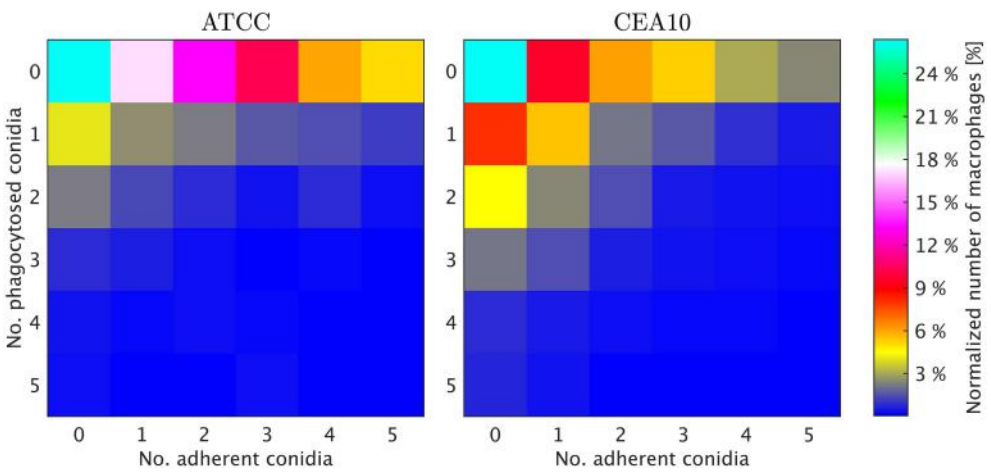


FIGURE 6 | Two-dimensional probability distributions of adherent and phagocytosed conidia over macrophages for the two *A. fumiatus* strains ATCC and CEA10.

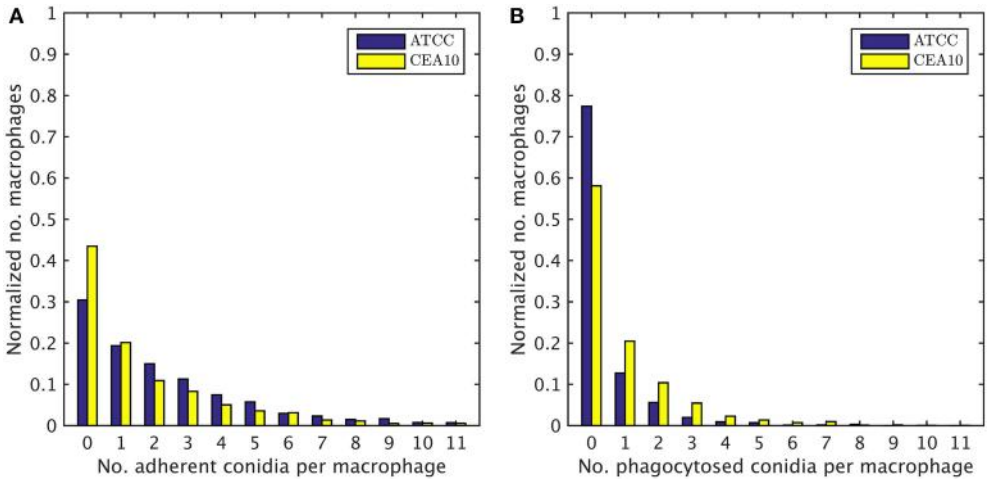


FIGURE 7 | One-dimensional probability distributions for a macrophage to have a certain number of (A) adherent and (B) phagocytosed conidia.

$$\varphi_i = \frac{N_c^{phag}}{N_m} \cdot \varphi_m,$$

corresponds to the product of the number of phagocytosed conidia per macrophage and the uptake ratio (Sano et al., 2003). **Figure 8C** shows the phagocytic index of both strains. Similarly, it could be argued for combining the uptake ratio with the phagocytosis ratio,

$$\varphi_i^{sym} = \varphi_c \cdot \varphi_m,$$

where the number of macrophage-associated conidia and the number of macrophages entercontribute in a more symmetric fashion and which is therefore referred to as symmetrized phagocytic index. This measure is presented in **Figure 8D** for comparison with the other measures. All measures and distributions point toward a higher degree of phagocytosis for the CEA10 strain compared to ATCC.

Aggregation Observed for Adherent But Not for Non-adherent Conidia

Rather than occurring as isolated cells, conidia were often observed to cluster in aggregates. We consider image objects to be aggregated if they have common borders. As explained in subsection Classification of the Materials and Methods section in the context of macrophage-adherence by a conidium, we computed common borders between conidia to identify such clusters. The aggregation ratio is defined by,

$$\gamma_r = \frac{N_c^{agg}}{N_c^{non-phag}},$$

where N_c^{agg} denotes all non-phagocytosed conidia which are aggregated and $N_c^{non-phag}$ is the number of all conidia which are non-phagocytosed. Thus, we did not account for phagocytosed conidia, because their visual aggregation might solely be appearing due to spatial constraints in the macrophage.

Since non-phagocytosed conidia could be adherent or non-adherent, we distinguished between the aggregation ratio for non-adherent conidia and for adherent conidia. For this purpose, we distinguished between adherent clusters and non-adherent clusters and we applied the same formula. In **Figure 9A** we observe a higher aggregation ratio for ATCC compared to CAE10 when considering all non-phagocytosed conidia. However, when we divided this population into adherent and non-adherent conidia (see **Figures 9B,C**) it was revealed that the difference in aggregation between the two strains was only present in the adherent conidia. As we have already demonstrated that ATCC conidia were adherent to macrophages to a higher degree than CEA10, whereas the latter is more likely to be phagocytosed, it may be argued that the difference in aggregation was merely the consequence of spatial limitation on the surface of macrophages. For the population of conidia that are neither phagocytosed nor adherent to a macrophage, both strains showed a very low aggregation ratios with no significant difference between them.

Discussion

Confrontation assays are commonly used today to quantify host-pathogen interactions between immune cells and pathogenic cells. In contrast to techniques based on flow cytometry, approaches based on microscopy images do provide a richer amount of information, e.g., on spatial correlations and morphological properties of cells. However, the information gained by image-based approaches is typically foiled by a tedious and error-prone manual analysis of the data. In this work, we addressed this drawback and developed a computer algorithm that performs the image analysis automatically and that opens up the possibility for high-throughput screening while retaining the full information content of the images. In fact, for the microscopy images analyzed in the present study, we achieved computation times per image of about 1 min, which was more than one order of magnitude lower than was required for a manual analysis.

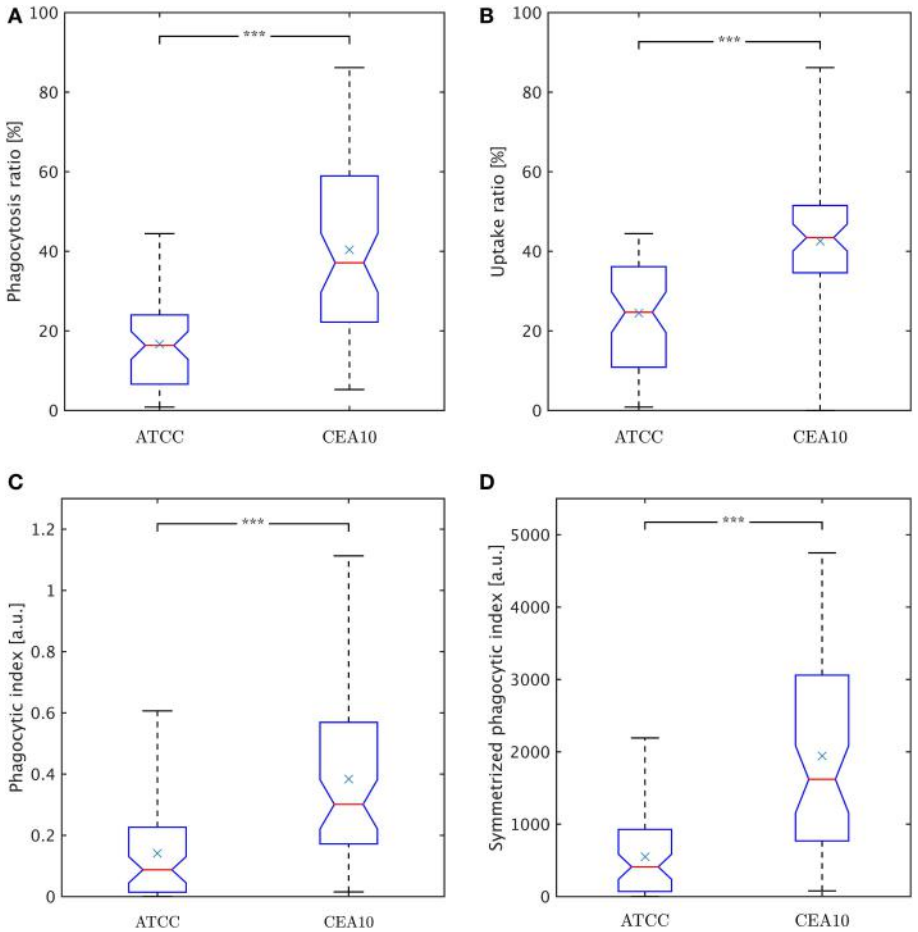


FIGURE 8 | Comparison of phagocytosis measures between the two *A. fumiatus* strains ATCC and CEA10. (A) Phagocytosis ratio φ_c. **(B)** Uptake ratio φ_m. **(C)** Phagocytic index φ_i. **(D)** Symmetrized phagocytic index φ_i^{sym}. See the text for details.

We performed a rigorous validation of the algorithm and obtained high performance measures for the overall sensitivity (96.6%), precision (97.8%), and accuracy (94.5%), which are reasonable values in the light of the high variation of image objects. We identified the main source of errors to be the similarity in shape and intensity properties between image objects and background regions. For example, this concerned background regions that were completely surrounded by macrophages or conidia, or macrophages that were of too low signal as a result of insufficient staining, or clusters with slightly superimposed conidia that could not be correctly identified because of their effectively smaller cluster size.

Algorithms for the automated image analysis of confrontation assays, especially in the context of fungal pathogens interacting with immune cells, have been developed before (Mech et al., 2011, 2014; Kraibooj et al., 2014; Schäfer et al., 2014). The most important progress of the novel algorithm presented here concerns the successful segmentation of macrophages. This was not achieved previously, because it was complicated by the occasionally interrupted staining of the macrophage surface, but is a necessary prerequisite for the comprehensive quantification of these phagocytosis assays. Thus, having achieved this crucial progress in the present work, the task of quantifying this type of phagocytosis assays was now comprehensively solved, because spatial and functional information on the interaction between all cells in the assay is now accessible. For example, we demonstrated that macrophages involved in the phagocytosis process can be detected and whether macrophage-associated conidia were really phagocytosed or just adherent. This enables building up a two-dimensional distribution of conidia over the macrophages from which other measures describing the immune response can be derived.

Interestingly, different measures have been proposed for the quantification of biological processes such as phagocytosis events. Taking either the viewpoint of the pathogens or the immune cells, we here computed the phagocytosis ratio φ_c (Mech et al., 2011; Kraibooj et al., 2014) and the uptake ratio φ_m (Sano et al., 2003), respectively. A quantity that combines both viewpoints is commonly referred to as phagocytic index φ_i (Sano et al., 2003), for which a symmetrized measure $\varphi_i^{sym} = \varphi_c \cdot \varphi_m$ was proposed here. Obviously, all these measures will provide different absolute numbers and can be used in the comparison of different strains.

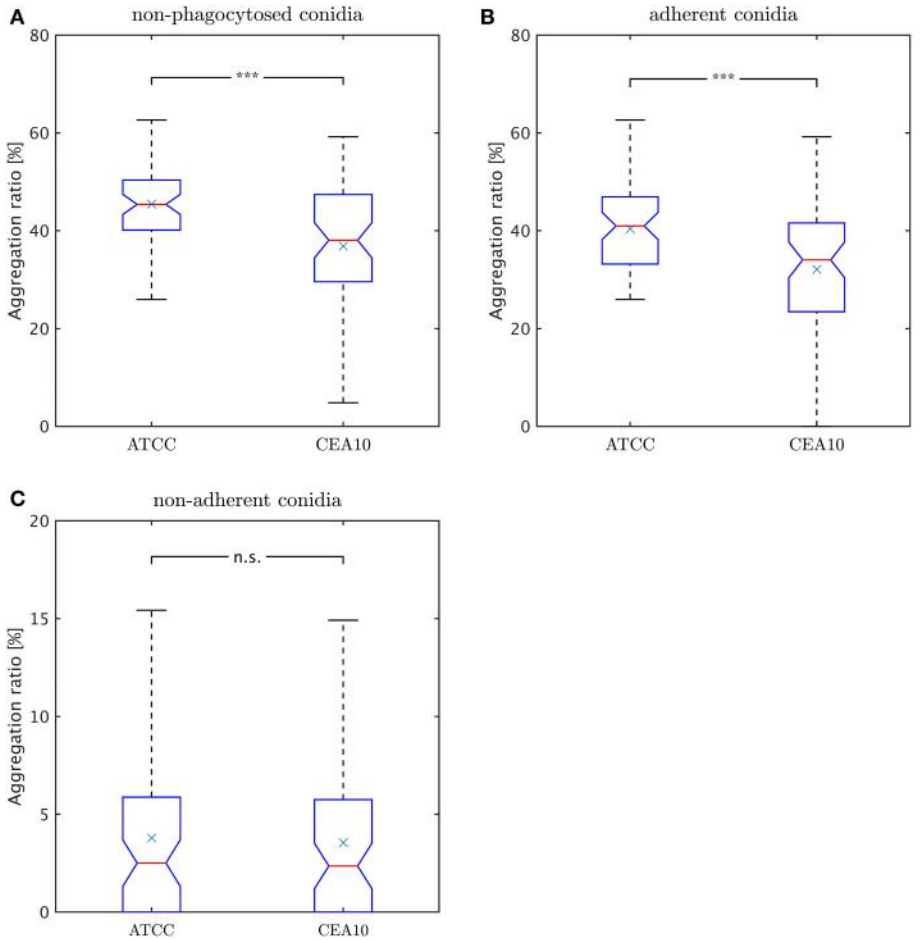


FIGURE 9 | Comparison of aggregation ratio, γ_r, of non-phagocytosed conidia between the two *A. fumiatus* strains ATCC and CEA10. Aggregation ratio relative to **(A)** all non-phagocytosed conidia, **(B)** adherent conidia, and **(C)** non-adherent conidia.

More importantly, they have to be interpreted with some care, which is in particular true for the measures with the combined viewpoint. This is a consequence of the fact that these measures involve a multiplication of two factors. For example, as can be easily demonstrated for the symmetrized phagocytic index, a high value of the phagocytosis ratio φ_c and a low value of the uptake ratio φ_m yield a value for φ_i^{sym} that can be identical for low φ_c and high φ_m and by that masks important differences in the underlying biology. Therefore, we conclude that, rather than representing the immune response by a single scalar measure, interpretation of the confrontation assay should be inferred from a comparison of the distribution of pathogens over immune cells. If the confrontation assays contain more than two different cell types, the distribution can be extended to higher dimensions as long as sufficient data are available.

In the present study, conidia of strain CEA10 showed a significantly higher phagocytosis ratio than ATCC. This was not only confirmed for the viewpoint of conidia by the phagocytosis ratio and of macrophages by the uptake ratio, but also by the combined viewpoint of conidia and macrophages in terms of the phagocytic index and the symmetrized phagocytic index. From the quantities of these measures, it can be concluded that the experiments were neither limited by the saturation of macrophages with phagocytosed conidia nor by the depletion of non-phagocytosed conidia. Thus, the significant differences in all measures indicated that the process of conidia recognition and uptake was generally more effective for CEA10 than for ATCC. In particular, this could be concluded from the fact that not only the percentage of phagocytosed conidia was higher for CEA10 than for ATCC, but also the percentage of phagocytosing macrophages. The conclusion that the initiation of phagocytosis was less effective for ATCC was also in line with the observation that the number of ATCC conidia that were adherent to macrophages were higher compared to CEA10 conidia. Furthermore, we could exclude that differences in the phagocytosis of conidia were due to the aggregation of conidia, because no significant differences in the aggregation ratio were observed for the non-adherent conidia of the two strains.

Previous studies on an embryonated egg model and in mice showed that CEA10 was more virulent than ATCC (Jacobsen

et al., 2010; Heinekamp and Brakhage, 2012). The question remains if the difference in virulence is directly correlated to the observed difference in phagocytosis ratio. In a similar study on *Lichtheimia corymbifera* it was also shown that a more virulent strain was more effectively phagocytosed (Kraibooj et al., 2014). In case the spores are able to inhibit killing after being phagocytosed they could use macrophages as a survival niche and escape from the phagocyte by germination (Amin et al., 2014). However, further experiments would have to be performed to prove this hypothesis.

We expect that our novel algorithm for fully automated image analysis will be of importance in future research for several reasons. Firstly, it paves the way for comparative high-throughput screening of mutant collections and their comprehensive quantification. Secondly, the algorithm is generally applicable to assays of cells with close-to circular morphology and can be straightforwardly extended to assays for more than two different cell types. Thirdly, the results achieved here form a quantitative data base for the development of mathematical models that enable realistic simulations of biological processes on the computer. Image-based systems biology is a modern field of research (Medyukhina et al., 2015) and has a plethora of applications, for example, providing image-derived techniques for differentiating between cell colocalization and random positioning of cells (Mokhtari et al., 2015) or simulating virtual infection models for *A. fumigatus* infection (Tokarski et al., 2012; Pollmächer and Figge, 2014).

Author Contributions

Conception and design of the investigation and work: CS, AB, MTF. Contribution of materials and computational resources: AB, MTF. Imaging experiments: HS. Data processing, development and application of the computer algorithm: KK, CS. Evaluation and analysis of the results: KK, HS, CS, AB, MTF. Drafting the manuscript and revising it critically for important intellectual content and final approval of the version to be published: KK, HS, CS, AB, MTF. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: KK, HS, CS, AB, MTF.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) within CRC/TR 124 *Human-pathogenic fungi and their human host—networks of interaction* FungiNet (project A1 to AB and project B4 to MTF). The funder has no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00549/abstract

References

Aderem, A., and Underhill, D. M. (1999). Mechanisms of phagocytosis in macrophages. *Annu. Rev. Immunol.* 17, 593–623. doi: 10.1146/annurev.immunol.17.1.593

Amin, S., Thywissen, A., Heinekamp, T., Saluz, H. P., and Brakhage, A. A. (2014). Melanin dependent survival of Apergillus fumigatus conidia in lung epithelial cells. *Int. J. Med. Microbiol.* 304, 626–636. doi: 10.1016/j.ijmm.2014.04.009

Blinchikoff, H., and Krause, H. (2001). *Filtering in the Time and Frequency Domains*. London: The Institution of Engineering and Technology.

Brakhage, A. A. (2005). Systemic fungal infections caused by Aspergillus species: epidemiology, infection process and virulence determinants. *Curr. Drug. Targets.* 6, 875–886. doi: 10.2174/138945005774912717

Brakhage, A. A., Jahn, B., and Schmidt, A. (eds.). (1999). *Aspergillus Fumigatus: Biology, Clinical Aspects, and Molecular Approaches to Pathogenicity*, Vol. 2. Basel: Karger Medical and Scientific Publishers.

Brakhage, A. A., and Langfelder, K. (2002). Menacing mold: the molecular biology of *Aspergillus fumigatus*. *Annu. Rev. Microbiol.* 56, 433–455. doi: 10.1146/annurev.micro.56.012302.160625

Brakhage, A. A., and Van den Brulle, J. (1995). Use of reporter genes to identify recessive trans-acting mutations specifically involved in the regulation of *Aspergillus nidulans* penicillin biosynthesis genes. *J. Bacteriol.* 177, 2781–2788

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7:R100. doi: 10.1186/gb-2006-7-10-r100

Chambers, J. M. (1983). *Graphical Methods for Data Analysis*. Belmont: Wadsworth International Press.

Dagenais, T. R., and Keller, N. P. (2009). Pathogenesis of *Aspergillus fumigatus* in invasive aspergillosis. *Clin. Microbiol. Rev.* 22, 447–465. doi: 10.1128/CMR.00055-08

Fabbri, R., Costa, L. D. F., Torelli, J. C., and Bruno, O. M. (2008). 2D Euclidean distance transform algorithms: a comparative survey. *ACM Comput. Surv.* 40, 2. doi: 10.1145/1322432.1322434

Gonzalez, R. C., and Woods, R. E. (2006). *Digital Image Processing, 3rd Edn*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Heinekamp, T., and Brakhage, A. A. (2012). "Genome plasticity of aspergillus species," in *Genome Plasticity and Infectious Diseases*, eds J. Hacker, U. Dobrindt, and R. Kurth (Washington, DC: American Society of Microbiology), 326–342.

Jacobsen, I. D., Große, K., Slesiona, S., Hube, B., Berndt, A., and Brock, M. (2010). Embryonated eggs as an alternative infection model to investigate *Aspergillus fumigatus* virulence. *Infect. Immun.* 78, 2995–3006. doi: 10.1128/IAI.00268-10

Kraibooj, K., Park, H. R., Dahse, H. M., Skerka, C., Voigt, K., and Figge, M. T. (2014). Virulent strain of *Lichtheimia corymbifera* shows increased phagocytosis by macrophages as revealed by automated microscopy image analysis. *Mycoses* 57, 56–66. doi: 10.1111/myc.12237

Krzywinski, M., and Altman, N. (2014). Points of significance: visualizing samples with box plots. *Nat. Methods* 11, 119–120. doi: 10.1038/nmeth.2813

Latgé, J. P. (1999). *Aspergillus fumigatus* and aspergillosis. *Clin. Microbiol. Rev.* 12, 310–350.

Maerker, C., Rohde, M., Brakhage, A. A., and Brock, M. (2005). Methylcitrate synthase from *Aspergillus fumigatus*. Propionyl-CoA affects polyketide synthesis, growth and morphology of conidia. *FEBS J.* 272, 3615–3630. doi: 10.1111/j.1742-4658.2005.04784.x

Mech, F., Thywißen, A., Guthke, R., Brakhage, A. A., and Figge, M. T. (2011). Automated image analysis of the host-pathogen interaction between phagocytes and *Aspergillus fumigatus*. *PLoS ONE* 6:e19591. doi: 10.1371/journal.pone.0019591

Mech, F., Wilson, D., Lehnert, T., Hube, B., and Figge, M. T. (2014). Epithelial invasion outcompetes hypha development during *Candida albicans* infection

as revealed by an image-based systems biology approach. *Cytometry A* 85, 126–139. doi: 10.1002/cyto.a.22418

Medyukhina, A., Timme, S., Mokhtari, Z., and Figge, M. T. (2015). Image-based systems biology of infection. *Cytometry A* 87, 462–470. doi: 10.1002/cyto.a.22638

Mokhtari, Z., Mech, F., Zehentmeier, S., Hauser, A. E., and Figge, M. T. (2015). Quantitative image analysis of cell colocalization in murine bone marrow. *Cytometry A* 87, 503–512. doi: 10.1002/cyto.a.22641

O'Gorman, C. M., Fuller, H. T., and Dyer, P. S. (2008). Discovery of a sexual cycle in the opportunistic fungal pathogen *Asperigillus fumigatus*. *Nature* 457, 471–474. doi: 10.1038/nature07528

Pal, N. R., and Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognit.* 26, 1277–1294. doi: 10.1016/0031-3203(93)90135-J

Pollmächer, J., and Figge, M. T. (2014). Agent-based model of human alveoli predicts chemotactic signaling by epithelial cells during early *Aspergillus fumigatus* infection. *PLoS ONE* 9:e111630. doi: 10.1371/journal.pone.0111630

Rittscher, J. (2010). Characterization of biological processes through automated image analysis. *Annu. Rev. Biomed. Eng.* 12, 315–344. doi: 10.1146/annurev-bioeng-070909-105235

Roerdink, J. B. T. M., and Meijster, A. (2001). The watershed transform: definitions, algorithms and parallelization strategies. *Fund. Inform.* 41, 187–228. Available online at: http://www.cs.rug.nl/roe/publications/parwshed.pdf

Sano, H., Hsu, D. K., Apgar, J. R., Yu, L., Sharma, B. B., Kuwabara, I., et al. (2003). Critical role of galectin-3 in phagocytosis by macrophages, *J. Clin. Invest.* 112, 389–397. doi: 10.1172/JCI200317592

Schäfer, K., Bain, J. M., Di Pietro, A., Gow, N. A., and Erwig, L. P. (2014). Hyphal growth of phagocytosed *Fusarium oxysporum* causes cell lysis and death of murine macrophages. *PLoS ONE* 9:e101999. doi: 10.1371/journal.pone.0101999

Schönmeyer, R., Athelogou, M., Sittek, H., Ellenberg, P., Feehan, O., Schmidt, G., et al. (2011). Cognition Network Technology prototype of a CAD system for mammography to assist radiologists by finding similar cases in a reference database. *Int. J. Comput. Assist. Radiol. Surg.* 6, 127–134. doi: 10.1007/s11548-010-0486-8

Sezgin, M., and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *J. Electron. Imaging.* 13, 146–168. doi: 10.1117/1.1631315

Thywißen, A., Heinekamp, T., Dahse, H. M., Schmaler-Ripcke, J., Nietzsche, S., Zipfel, P. F., et al. (2011). Conidial Dihydroxynaphthalene Melanin of the Human Pathogenic Fungus *Aspergillus fumigatus* Interferes with the Host Endocytosis Pathway. *Front. Microbiol.* 3:96. doi: 10.3389/fmicb.2011.00096

Tokarski, C., Hummert, S., Mech, F., Figge, M. T., Germerodt, S., Schroeter, A., et al. (2012). Agent-based modeling approach of immune defense against spores of opportunistic human pathogenic fungi. *Front. Microbiol.* 3:129. doi: 10.3389/fmicb.2012.00129

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bull.* 1, 80–83. doi: 10.2307/3001968

Zhou, X., and Wong, S. T. (2006). Informatics challenges of high-throughput microscopy. *IEEE Signal Process. Mag.* 23, 63–72. doi: 10.1109/MSP.2006.1628879

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on