MULTISCALE ANALYSIS *of* INFECTIOUS DISEASES

INTEGRATING OMICS AND CLINICAL INFORMATICS DATA INTO PATIENT CARE

Theodore Robertson Pak

April 27, 2017

A dissertation submitted to the Graduate Faculty
of the Graduate School of Biomedical Sciences,
Biomedical Sciences Doctoral Program,

in partial fulfillment of the requirements
for the degree of Doctor of Philosophy,
Icahn School of Medicine at Mount Sinai.

This manuscript has been read and accepted by the Graduate Faculty of the Graduate School of Biomedical Sciences, in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Dissertation Advisor — Andrew Kasarskis, PhD ..... Jun 29, 2017
Date

Dean — Marta Filizola, PhD 6/26/17
Date

DISSERTATION COMMITTEE

James Iatridis, PhD (chair)
Adolfo García-Sastre, PhD
Joel Dudley, PhD
Jonathan Karr, PhD
Jun Zhu, PhD
Bo Shopsin, MD, PhD (outside reviewer)

ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI
2017

Abstract

Multiscale analysis of infectious diseases:
Integrating omics and clinical informatics data into patient care

by Theodore Robertson Pak

Advisor: Andrew Kasarskis, PhD

New sources of data, such as next generation sequencing (NGS) of pathogen genomes, electronic medical records (EMR), and omics assays like RNA-seq and mass cytometry, are poised to transform clinical infectious diseases. One of the greatest challenges in applying these "big data" toward clinical practice is the development of bioinformatics techniques and software that make downstream analyses routine, rigorous, and actionable. Herein, I develop software and integrative statistical models for several combinations of these data to address urgent global threats in infectious diseases, including the spread of healthcare associated infections (HAIs), skyrocketing rates of antimicrobial resistance, and understanding human immune responses to poorly characterized mosquito-borne viruses.

I first demonstrate that *de novo* assembly of long reads can finish the genomes of hospital bacterial isolates with resolution sufficient for discovering a resistance-conferring single-nucleotide variant that emerges during failed antimicrobial therapy—in our case, a quinolone resistance variant in a strain of *Stenotrophomonas maltophilia*. I then describe two software packages, a new genome browser and a suite of modular bioinformatics pipelines, that facilitate the routine usage of NGS by the Pathogen Surveillance Program at The Mount Sinai Hospital for reconstructing HAI transmission networks between patients. These software additionally provide actionable visualizations of genomic data tailored to infection control physi-

cians. The use of long read data for hospital surveillance is novel and offers unique opportunities to capture recombination and horizontal transfer events occurring throughout the rapid evolution of virulence and resistance in HAI strains.

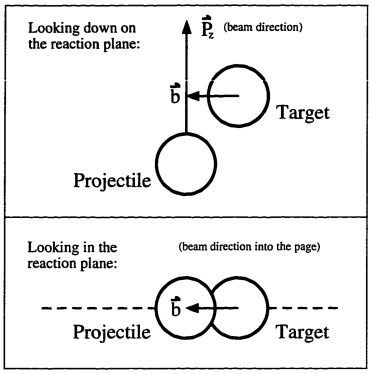
To convert these data into management strategies, I use machine learning algorithms on seven years of EMR data at Mount Sinai to precisely estimate the marginal cost per patient of *Clostridium difficile* infection, helping to anchor a cost-benefit analysis for selecting interventions and surveillance investments. Finally, in a more futuristic study, I integrate RNA-seq, mass cytometry, and multiplexed immunoassay data to comprehensively profile the human immune response to chikungunya virus, a re-emerging arthritogenic arbovirus, which reveals a significant role for monocytes and novel biomarkers for clinical outcomes like symptom severity and immunogenicity. Because of the many applications for multiscale analysis I establish in this dissertation, I conclude that it is indeed transforming infectious diseases.

Acknowledgments

THE LIST OF PEOPLE whom I must thank for helping me attain this doctorate degree is quite long, and there is no simple way for me to adequately express my gratitude to each and every one of them. A significant part of this PhD has been realizing how greatly my education and research have depended on the time, sweat, trust, and patience of so many others, who by God's grace decided to invest in my future. What follows is my best attempt at acknowledging those that gave me so much.

TO MY MOTHER AND FATHER, Dr. Robert and Myung-Hee Pak. My father, who was the first PhD in his own family, whom I remember happily marching down the aisles of the Breslin Center to the honks of Pomp and Circumstance for his doctorate in nuclear physics. Dad, you've been the role model for my life. Ever since I helped draw two figures for your thesis on the Macintosh you got me as a kid, proudly feeling like I had just put the torch on the Statue of Liberty, I've been wondering what it would be like to write one of my own. Well, have I got some bedtime reading material for you! And my mother, who kept me alive and kicking from a single cell all the way to whatever I am today, which if you think about it, is more remarkable than anything you will read in this dissertation. This never could have been written had it not been for my mother's ferocious dedication toward making a better life for me. Thanks, ma!

TO MY THESIS ADVISOR, Dr. Andrew Kasarskis, for taking me under his wing for the past four years. I am supremely lucky to have met somebody with the rare blend of patience, humor, genuine scientific curiosity, and mentorship skills that Andrew has and shares with others so freely. Andrew has taught me more than just science—he has taught me about leadership, entrepreneurship, integrity, life, and occassionally, wildlife. I look forward to pondering his lessons for the rest of my career. And to my fellow Kasarskis brother from the halls of Regis, the inimitable Dr. Joseph R. Scarpa, who blazed a trail for us both and kept me company as we jockeyed desks in Greenland. I am fortunate to be returning to medical school at his side.



One of the figures I helped my dad draw on a computer for his thesis, "Collective flow in intermediate energy heavy-ion collisions," in 1996. Perhaps one day, I too will understand what it means.

To the boys of 7G, who are back in town, spread the word around. Kevin "Carpet" Hoffman, Michael "The Stuffer" Daniel, Zachary "BJ" Lorsch, Eddie "Other Ted" Contijoch, and Ranjan "Gammie" Upadhyay were my roommates as we began this strange and imprudent quest toward a dual degree. Then there's Andrew T. McKenzie, a late but profitable addition to the gang, who singlehandedly doubled our weirdness quotient and may just achieve his goal of a completely rational life strategy within the decade. And my entire entering class of 2012, who shall be rightfully remembered as "the Titans." We have been through much together: retreats, pyramids, marathons, camping trips, uninhibited human flight, weddings, even a Tough Mudder—many of which could be considered a microcosm of certain aspects of the MSTP experience, but all of which I will always remember. You all kept me variously motivated, alive, and inspired throughout this ignoble quest, and I'll enjoy seeing you all crush it at science, medicine, and life.

To the entire Sinai MSTP, starting with the director from when I entered, Dr. Yasmin Hurd, and the many other leaders that have guided me since: Dr. Margaret Baron, Dr. Talia Swartz, Dr. Benjamin Chen, and Dr. Scott Friedman, to name a few. To the many other students in the program that gave me guidance and support, not least of all Dr. Benjamin Laitman, who was one of the first Sinai students I met, who steadfastly supported me through the challenges of leading Student Council, and whom I am honored to call my friend.

To my colleagues in science, who are also listed in acknowledgements within each chapter, but whom I must thank here not only for their collaborative efforts but also their friendship, mentorship, and personal support. From the Icahn Institute and Department for Genomics and Multiscale Biology: El-Ad David Amir, Oliver Attie, Ali Bashir, Harm van Bakel, Kieran Chacko, Brianne Ciferri, Gintaras Deikus, Gang Fang, Zeynep Gumus, Seunghee Kim-Schulze, Martha Lewis, David Nathanson, Leah C. Newman, Tim O'Donnell, Adeeb Rahman, Eric Schadt, Erick Scott, Robert Sebra, Mayte Suarez-Fariñas, Mitchell Sullivan, Maria Suprun, and Elizabeth Webster. From the Departments of Medicine and Pathology of the Mount Sinai Hospital: Judith Aberg, Camille Hamula, Jonathan Hand, Shirish Huprikar, Gopi Patel, Timothy Sullivan, and Fran Wallach. From the Department of Microbiology of the Icahn School of Medicine at Mount Sinai: Ana Fernandez-Sesma, Rebecca Hamlin, and Irene Ramos-Lopez, aka the Divas of Virology, who were so amazing to work with. To my colleagues from other institutions in the Dengue Human Immune Profiling Consortium, including Eva Harris and Daniela Michlmayr from UC Berkeley and Steven Wolinsky and Eun-Young Kim from Northwestern. There are likely many more scientists that quietly contributed their technical insight and effort to some aspect of the protocols, techniques, and datasets that eventually became a part of this dissertation, and for all of these unsung heroes whom I may never even meet, I want to express my sincere gratitude.

There are many other unsung heroes in and around Mount Sinai whose support has been invaluable. Firstly, Courtney Manning, Gayle Schneiderman, and Rhaisili Rosario, who all did gangbusters work in keeping the MSTP running. Dr. Rainier P. Soriano, Dr. Yasmin Meah, Dr. David C. Thomas, Paul Lawrence, and Dr. David Muller gave me crucial guidance at certain points in the journey. I was lucky to have advice from Geoffrey Smith and Dan Seltzer on starting my business, The East Harlem Software Company, Inc. I have to thank the staff of Aron Hall, for maintaining such a nice home for all of us students. The guys at El Aguila on 103rd and Lex, whose tortas Cubanas fueled the improbable appearance of many words on these pages (which I recently learned is not an actual food in Cuba, but such is America). No acknowledgements could be complete without saluting Andy Efros, literally the nicest guy at Mount Sinai.

I especially want to thank Dr. Deena Altman for her dedicated leadership and evangelism of the Pathogen Surveillance Program, whose clinical insight and contributions made much of this dissertation possible before I even began, and who supported me from the first day I joined the group. As I go back into the hospital for clerkships and wonder what kind of doctor I want to be, I will be thinking often of Deena.

To the members of my thesis advisory committee: Dr. Adolfo García-Sastre, Dr. Jun Zhu, Dr. Joel Dudley, Dr. Jonathan Karr, and the chair Dr. James Iatridis, for providing focused guidance in developing this dissertation and meticulously supporting my development as a scientist. Special thanks to Dr. Bo Shopsin from the NYU School of Medicine, who graciously agreed to be an outside reviewer. To my undergraduate mentor, Dr. Frederick "Fritz" Roth, and members of his laboratory, who supported me throughout and after college, and first inspired me to pursue a doctorate in bioinformatics.

Finally, I'd like to thank Dr. Sonia Yen Jarrett, the only lady brave enough to put up with my shenanigans, who has kept me motivated throughout challenges I once thought impossible, and who always makes me laugh.

Contents

List of Tables

List of Figures

1

Introducing next-generation sequencing and multiscale data analysis into clinical infectious diseases

Recent reviews have suggested that routine next-generation sequencing (NGS) on clinical specimens will improve the capabilities of clinical microbiology laboratories. However, the real opportunity to impact our understanding and management of infectious diseases lies in integrating NGS with clinical data from electronic medical records (EMRs), immune profiling data, and other rich datasets to create multiscale predictive models. This chapter introduces a range of new "omics" and patient data sources relevant to infectious diseases and proposes three potentially disruptive applications for these data in the clinical workflow. The combined threats of healthcare-associated infections and multidrug resistant organisms may be addressed by multiscale analysis of NGS and EMR data that is ideally updated and refined over time within each healthcare organization. Such data and analysis should form the cornerstone of future learning health systems for infectious disease.

What is the most resilient parasite? Bacteria? A virus? An intestinal worm? An idea. *Resilient... highly contagious. Once an idea has taken hold of the brain it's almost impossible to eradicate.*

—COBB, *Inception*

C-3PO: *Sir, it's quite possible this asteroid is not entirely stable.*

HAN SOLO: *Not entirely stable. I'm glad you're here to tell us these things. Chewie! Take the Professor in back and plug him into the hyperdrive!*

—Star Wars: Ep. V – The Empire Strikes Back

NEXT-GENERATION SEQUENCING and analysis techniques for "big data" are poised to transform our understanding of diseases that have a complex inherited component, such as cancer, diabetes, and heart failure. Perhaps even more significant, however, is the impact these technologies will have on the management of *infectious diseases*, which have discrete, identifiable causes that can be isolated, cultured, and tested against drugs in vitro as part of a standard clinical workflow. Despite steady technological improvements in each step, this workflow's principles have not changed for a century.[1]

Our capacity to acquire "omics" data about infections is increasing exponentially. Nanoscale parallelization of DNA sequencing has precipitously dropped the cost per base-pair of finished genomes while increasing throughput, and

[1] Didelot et al. (2012), "Transforming clinical microbiology with bacterial genome sequencing"; Köser et al. (2012), "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology."

the cost of sequencing and assembling a bacterial genome is trending below $100.[2] PacBio RS sequencing has increased median read lengths over 10kbp, facilitating rapid, automated finishing of genomes for outbreak pathogens.[3] Beyond sequencing pathogen genomes, recent studies have used other "omics" experimental techniques such as Luminex cytokine assays, RNA-seq, and mass cytometry to characterize immune responses to infection or vaccination with remarkable precision.[4]

Many public databases curate and disseminate "omics" data relevant to infectious disease (Table 1.1), but most lack significant clinical metadata. Increasing adoption of electronic medical records (EMRs) can potentially mitigate this problem because they typically include data on demographics, medications, lab results, and more. Figure 1.1 presents a visualization of some of these datatypes as recorded by Mount Sinai Hospital's EMR for two patients diagnosed with *Clostridium difficile* colitis. With so many different stakeholders entering EMR data, however, automatically extracting certain facts (e.g., "this patient had the flu last Tuesday") can be difficult. Nevertheless, high-accuracy methods for extracting infectious phenotypes such as influenza-like illness [5], unclear HIV status,[6] and community-acquired pneumonia[7] have been demonstrated, and consortia such as eMERGE are standardizing comparison, validation, and deposition of these algorithms into a central repository.[8]

The marriage of real-time digital clinical information with "omics" technology creates the opportunity to increase the precision of clinical decision-making

[2] Didelot et al. (2012).

[3] Chin et al. (2011), "The Origin of the Haitian Cholera Outbreak Strain"; Rasko et al. (2011), "Origins of the E. Coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany".

[4] Mejias and Ramilo (2014), "Transcriptional profiling in infectious diseases: ready for prime time?"; Querec et al. (2009), "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans."

Database focus	For general research	For infectious disease	
		Multi-pathogen	Pathogen-specific
Genomes	• NCBI Nucleotide (GenBank/RefSeq) • ENA/EMBL • DDBJ	• ViPR • NMPDR • PATRIC • EuPathDB	• Influenza Research Database (IRD) • Tuberculosis Database (TBDB) • LANL: Databases for HIV, HCV, and HFV
Gene products and functionality	• UniProt • KEGG	• Pathogen-Host Interaction Database • Antibiotic Resistance Genes Database • Comprehensive Antibiotic Resistance Database	
Expression and immune profiles	• GEO • ArrayExpress	• ImmPort	

Table 1.1: Examples of public bioinformatics databases that may be leveraged for multiscale analysis of infectious disease (this list is not exhaustive).

[5] Silva et al. (2013), "Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports"

[6] Felsen et al. (2014), "Development of an electronic medical record-based algorithm to identify patients with unknown HIV status."

[7] DeLisle et al. (2013), "Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy."

[8] Pathak, Kho, and Denny (2013), "Electronic health records-driven phenotyping: challenges, recent advances, and perspectives."

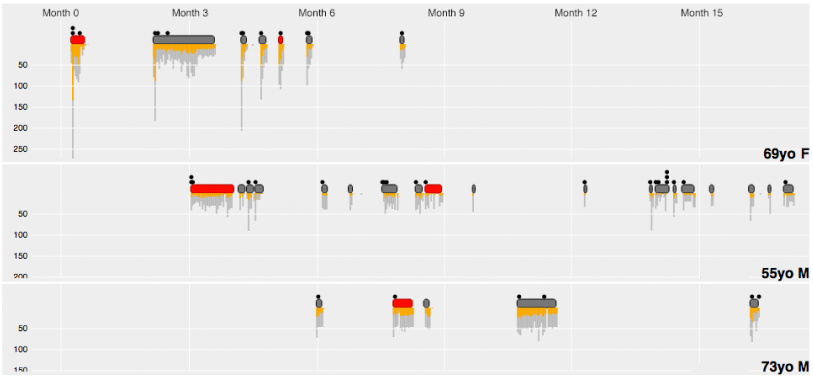


Figure 1.1: **Visualization of EMR data.** Timelines for two patients generated from EMR data in the Mount Sinai Data Warehouse (original analysis). Patient visits to Mount Sinai are represented as horizontal bars across the top of the timeline, and visits associated with a *C. difficile* infection are highlighted in red. The number of lab tests ordered per day is represented on the vertical axis, with abnormal results highlighted in orange. Transfer events are marked by the black dots above the horizontal bars.

and challenges us to quickly design and execute bioinformatics analyses. Predictive modeling of infectious disease that incorporates EMR data is still rare, although one recent study generated a social network for hospital acquired infection from EMR data using recorded contacts between patients and caretakers.[9] Another found that statistical analysis of EMR data produces risk factors for *C. difficile* infection that outperform models based only on medically recognized risks.[10] Likely because of the difficulty of integrating data across so many levels, no published studies have yet bridged predictive modeling on EMR data with pathogen genome sequences or other "omics" data from individual patients. Yet, for infectious disease, this is exactly what will fulfill the vision of a rapid-learning health system[11] that converts the informational byproducts of healthcare recorded by practitioners into evidence for future decision-making. While EMR data holds details of the clinical process and outcomes, "omics" data ties it back to pathophysiology and the precise strain and host-pathogen interactions present in each patient. Together, they can fuel a "learning engine" that integrates heterogeneous data into new clinical insights, interventions, and therapies. We will discuss how to leverage current bioinformatics software to build such an engine, and how this engine will be able to attack currently insurmountable problems in the field.

[9] Cusumano-Towner et al. (2013), "A social network of hospital acquired infection built from electronic medical record data."

[10] Wiens, Guttag, and Horvitz (2014), "A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions."

[11] Committee on the Learning Healthcare System in America and Institute of Medicine (2014), *Best care at lower cost: the path to continuously learning health care in America*; Kohane, Drazen, and Campion (2012), "A glimpse of the next 100 years in medicine."

The genomic clinical microbiology laboratory

Previous reviews[12] have proposed that cheap sequencing technology will transform clinical microbiology, while acknowledging technical and informa-

[12] Didelot et al. (2012); Köser et al. (2012).

tional barriers to adoption. Whole genome sequencing (WGS) via NGS provides ultimate resolution for epidemiological studies of transmission and relatedness, and may soon be cost-effective for routine use.[13] For pathogen identification, however, NGS is unlikely to usurp robotic culturing systems (e.g., Vitek and BD Phoenix) or newer mass spectrometry systems by cost and sensitivity comparisons alone, although it can lower turnaround time for difficult-to-culture organisms and identify novel or rarely-seen pathogens.[14] Since susceptibility or resistance of an organism to drugs is in principle fully encoded in its genetic material,[15] NGS can also lower turnaround times for drug susceptibility testing of slow-growing organisms, such as M. tuberculosis[16] and HIV-1.[17] This strategy should only expand as fuller catalogs of genomic variants that cause drug resistance are compiled for other pathogenic organisms.

[13] Didelot et al. (2012); Köser et al. (2012).

[14] Köser et al. (2012); Naccache et al. (2015), "Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult With Encephalitis by Unbiased Next-Generation Sequencing".

[15] Didelot et al. (2012); Gordon et al. (2014), "Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing."

[16] Boehme et al. (2010), "Rapid molecular detection of tuberculosis and rifampin resistance."

[17] Ram et al. (2015), "Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene population sequencing in the clinical HIV laboratory".

Leveraging existing bioinformatics tools

An oft-mentioned hurdle[18] for widespread use of NGS in clinical microbiology is the lack of readily accessible software for converting these data into species identifications, phylogenies, and drug susceptibilities. However, many mature open source bioinformatics solutions for individual components of these problems exist, and connecting these components into a pipeline is therefore a tractable software engineering exercise. Examples for most subtasks are listed in Table 1.2. As NGS use by clinical microbiology laboratories becomes more commonplace, we might anticipate full-fledged genomic clinical microbiology software packages to become widely available.

[18] Didelot et al. (2012); Köser et al. (2012).

This expectation has three foreseeable shortcomings. The first is that current tools are tied to centrally curated repositories of evidence. Although proponents of genomic clinical microbiology often envision encyclopedic databases hosted by international consortia,[19] human curation is expensive and inefficient at scale, and many infectious diseases are locale-specific phenomena. Models based on pooled data may fail to reflect variation between healthcare delivery regions;[20] for instance, a recent fitness model of H3N2 influenza based on international genomic surveillance data creates predictions only at the resolution of clades spanning multiple continents.[21] Since implementation of NGS in a healthcare institution's microbiology laboratory produces copious sequencing data not easily shared through public databases, institutions

[19] Didelot et al. (2012); Köser et al. (2012).

[20] Reis and Mandl (2003), "Integrating syndromic surveillance data across multiple locations: effects on outbreak detection performance."; Wiens, Guttag, and Horvitz (2014).

[21] Luksza and Lässig (2014), "A predictive fitness model for influenza."

should prepare to manage repositories of local evidence and predictive models that work specifically for them. Over time, as data exchange interfaces are developed, institutions could form consortia to generalize analyses, which is a strategy that has successfully increased the power of human genome-wide association studies.[22]

[22] Gottesman et al. (2013), "The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future."; Kohane, Drazen, and Campion (2012)

Table 1.2: Selected published bioinformatics software packages or databases that address specific steps of clinical microbiology tasks using NGS data (this list is not exhaustive). Well-established tools are available for many specific subtasks.

Problem domain	**Software or database**
Strain typing	• Multi-Locus Sequence Typing (MLST) database
De novo assembly from long reads	• Celera • Hierarchical Genome Assembly Process
Species identification	
From clonal sample	• NCBI BLAST • GenBank • Other databases in Table 1.1
From non-clonal sample	
Meta-assembly	• AMOS • MIRA • MetaVelvet
Clustering and species annotation	• MEGAN • MG-RAST
Maximum likelihood phylogeny trees	• BEAST • RAxML • ClonalFrame • ClonalOrigin
Whole genome alignment	
For SNP calling	• Mummer • Mugsy • Harvest
For structural variant calling	• Mauve
Gene annotation	
Bacterial	• Glimmer • RAST • prokka
Drug resistance in bacteria	• Resfinder • ARG-ANNOT • Mykrobe predictor
Other	• Influenza Virus Sequence Annotation Tool

A second shortcoming is that current pathogen annotation tools primarily make predictions using the simplistic criterion of sequence similarity. Machine

learning (ML) algorithms could eventually integrate a wider array of genotypic features extractable from pathogen genomes—variant calls, putative gene and motif annotations, and more—and train holistic models that predict phenotypes. A "top-down," integrative model predicting limited phenotypes from genotype for Mycoplasma genitalium is available;[23] top-down predictions of virulence, however, add the substantial complexity of host interactions. Therefore, genome-wide ML models of virulence have mostly been "bottom-up," blind to mechanistic knowledge, and oriented toward even smaller-genome pathogens with considerable genomic surveillance data. ML on viral sequence features has predicted more effective antiretroviral combinations for HIV,[24] genetic markers for host selectivity within families of viruses,[25] and optimal strain selection for H3N2 influenza vaccines.[26] In general, given the explosion in available data, significant untapped potential remains for ML-based models that predict virulence, transmissibility, and drug resistance from pathogen genotypes.

[23] Karr et al. (2012), "A whole-cell computational model predicts phenotype from genotype."

[24] Lengauer and Sing (2006), "Bioinformatics-assisted anti-HIV therapy."; Zazzi et al. (2012), "Predicting Response to Antiretroviral Treatment by Machine Learning: The EuResist Project".

[25] Raj et al. (2011), "Identifying hosts of families of viruses: a machine learning approach."

[26] Luksza and Lässig (2014).

The third shortcoming is that for many common pathogens, these models are still limited by the paucity of clinical metadata linked to sequenced pathogens. Pathogen phenotypes accessible directly from EMRs include prognostic variables, such as length of stay and disposition, and lab results, such as drug susceptibilities. Although lab information systems (LIS) typically do not forward non-clinical results (e.g., growth curves) to EMRs, data exported from the LIS can help define richer phenotypes. For some diseases, EMRs will contain lab results that directly reflect infection severity, e.g., viral load for HCV and HIV patients,[27] while other diseases will require more complex criteria.[28] Natural language processing of physician notes will facilitate the extraction of complex, high-accuracy clinical phenotypes from the EMR.[29] Routine NGS of specimens and EMR data on drugs prescribed and administered will enable ad-hoc studies crossing pathogen genotypes against interventions and outcomes. Richer characterization of particular host-pathogen encounters may be provided by immune and molecular profiling of selected patients, as well as animal experiments that establish individual pathogen genetic associations and molecular mechanisms. Biomarkers derived from such data[30] could enhance predictive models built on a zealous integration of NGS and EMR data.

[27] Norton and Naggie (2014), "The clinical management of HCV in the HIV-infected patient."

[28] DeLisle et al. (2013); Klompas et al. (2008), "Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance."; Silva et al. (2013).

[29] Liao et al. (2015), "Development of phenotype algorithms using electronic medical records and incorporating natural language processing"; Silva et al. (2013).

[30] Mejias and Ramilo (2014); Querec et al. (2009).

The growth of EMR phenotype information associated with pathogen

genomes will spur a new generation of pathogenicity and risk models based on genomic data. Ideally, these models can drive a "learning engine" that integrates heterogeneous input data from an encounter with an infected patient and predict outcomes for possible interventions. Predictions can be delivered to physicians via clinical decision support systems that complement EMR functions by suggesting relevant actions within a patient's electronic chart. The closing of the EMR–NGS–EMR loop (Figure 1.2) should be the ultimate goal of bioinformatics pipelines for genomic clinical microbiology, because this would maximize the utility of data created for clinical encounters, continuously turning yesterday's observations and outcomes into evidence for tomorrow's predictions.[31]

[31] Committee on the Learning Healthcare System in America and Institute of Medicine (2014); Kohane, Drazen, and Campion (2012)

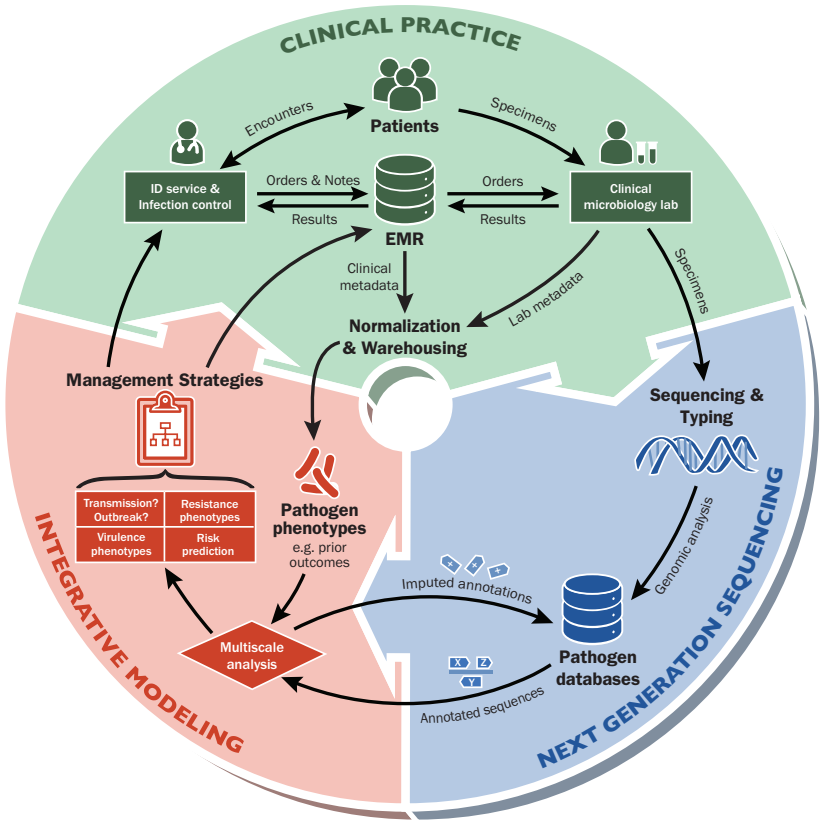


Figure 1.2: **A learning health system for infectious diseases.** Next generation sequencing (NGS) technologies now permit routine genomic analysis of clinical microbiology specimens. When integrated with pathogen phenotypes derived from clinical metadata in electronic medical records (EMRs) and laboratory metadata, we can generate predictive models for pathogen transmission, outbreaks, drug resistance, virulence, and risk factors for infection or critical outcomes that are specific to the health system and its patient population. If management strategies are formulated from these predictions and sent to infectious disease (ID) physicians and hospital infection control, a continuous loop of data analysis, application, and model refinement is created.

This sounds ambitious, but we can look to analogous software designed as subcomponents of learning healthcare systems to anticipate likely costs and avenues for development. The i2b2 platform[32] and its counterpart SCILHS[33] are vendor-agnostic solutions for extracting and unifying data across EMRs for reuse in cohort design and robust meta-analysis. The eMERGE consortium

[32] Kohane, Drazen, and Campion (2012).

[33] Mandl et al. (2014), "Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture."

stimulated the creation of SHARPn for normalization and natural language processing of EMR data[34] and CLIPMERGE for automated pharmacogenomics alerts.[35] For these examples, working software was created after 1-5 years of development with $100k-$10M of annual public grant funding.[36] If the aforementioned open-source software is leveraged, an equal scale of public funding and collaboration among academic medical centers could make similar strides toward the proposal in Figure 1.2. A modular framework allowed i2b2 to expand in scope organically after initial release,[37] suggesting that successful strategies should first aim for simple but clinically useful tasks such as identifying species and transmissions while anticipating the addition of more complex analyses via plugins and community contributions. In short, a reasonable investment in scrupulous software engineering could produce the seeds of a learning health system for infectious disease within the decade.

[34] Rea et al. (2012), "Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project".

[35] Gottesman et al. (2013).

[36] Gottesman et al. (2013); Kohane, Drazen, and Campion (2012); Mandl et al. (2014); Rea et al. (2012).

[37] Kohane, Drazen, and Campion (2012); Mandl et al. (2014).

Beyond genomic data

While NGS rapidly moves toward routine use by clinical microbiology laboratories where it can be integrated with EMR data, other advances in data collection on infectious diseases create significant opportunities for predictive modeling that may eventually impact clinical practice. We now review these additional sources of data.

Immune profiling

Host response to infection is not merely the result of single gene modulations or amplification of specific cell types, but rather consists of complex interacting networks of RNA transcription, protein signaling and metabolism that impact cellular, tissue, and whole organism behaviors. The nature of the modulations that are specific to the host's particular system relative to the state of the system at the time of infection ultimately determines risk and severity of disease. Recent studies have used "omics" scale experimental techniques to provide groundbreaking insight into immune responses, including classification of acute respiratory infections in children,[38] predicting immunogenicity of a vaccine,[39] and characterizing effects of aging that gradually decrease vaccine efficacy.[40]

[38] Mejias and Ramilo (2014).

[39] Querec et al. (2009); Furman et al. (2013), "Apoptosis and other immune biomarkers predict influenza vaccine responsiveness."

[40] Poland et al. (2014), "A systems biology approach to the effect of aging, immunosenescence and vaccine response".

In each of these studies, researchers took an unbiased, hypothesis-free approach to their design and observed as many properties of the immune system as was experimentally feasible before, during, and after a perturbation, such as vaccine administration. Such properties include cytokine levels as measured by Luminex assays, global changes in gene expression within various leukocyte populations as measured by RNA-seq or microarrays, and changes in cell populations with various surface markers as observed by flow and mass cytometry. With sufficient sample size, patterns of biomarkers can be linked to various clinical outcomes, e.g., the host becoming immunogenic to an antigen.[41] These biomarkers can then be investigated further for their functional role or used in new assays for point-of-care diagnosis. For common presenting conditions, like acute febrile illness, where current diagnostic methods poorly distinguish between bacterial and viral disease, this capability would allow prompt selection of the most appropriate therapy.[42] For diseases like dengue that do not have accurate markers of immunization, these markers will be essential for development of a successful vaccine.[43]

[41] *A study that demonstrates the use of these three omic assays to find profiles for specific outcomes of Chikungunya infection is presented in Chapter 6 of this dissertation.*

[42] Mejias and Ramilo (2014).

[43] Mahalingam, Herring, and Halstead (2013), "Call to action for dengue vaccine failure."

The internet

Figure 1.3: **Geospatial analysis, then and now**. A, Excerpt from a map published by John Snow in 1855 depicting a cluster of cholera cases around a water pump on Broad Street in London, England. B, Retrospective analysis of geocoded tweets in New York City classified by the probability of representing actual influenza cases during the 2012-2013 flu season, from Nagar et al. (2014). The primary outbreak cluster was determined to be in Northern Brooklyn.

Patients can now be associated with a trove of digitized information that can be mined to better understand infectious disease. Frequent social contacts are often captured in address books, email inboxes, and social networks like Facebook. Patients discuss symptoms online, which have been captured from search engine queries[44] or public Twitter streams to find probable cases of pertussis, whooping cough, and influenza,[45] although these studies have known limita-

[44] Ginsberg et al. (2009), "Detecting influenza epidemics using search engine query data."

[45] Nagel et al. (2013), "The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets."

tions.[46] Often this data is attached to geospatial information, which can be used to construct spatiotemporal models of the disease that reveal clusters and the directionality of spread within a city environment,[47] providing in real time the same type of analysis that 19th century anesthesiologist John Snow meticulously compiled to identify a London water pump handle as the source of a cholera outbreak (see Figure 1.3).[48] Global human movement patterns are also captured in air travel network usage data, which has previously been combined with genomic surveillance data to predict transmission dynamics of H3N2 influenza.[49] Unifying such diverse data in clinically relevant models with actionable outputs remains a significant challenge. On the other hand, the advent of cloud computing and open-source data science platforms like Apache Hadoop, used by Facebook and Walmart to model complex consumer behavior, suggests that algorithms for large, heterogeneous datasets will become increasingly accessible to healthcare providers and life sciences researchers.[50]

[46] Lazer et al. (2014), "Big data. The parable of Google Flu: traps in big data analysis."

[47] Nagar et al. (2014), "A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives."

[48] Buechner, Constantine, and Gjelsvik (2004), "John Snow and the Broad Street pump: 150 years of epidemiology."; Snow (1855), *On the Mode of Communication of Cholera*.

[49] Lemey et al. (2014), "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2."

[50] Mohammed, Far, and Naugler (2014), "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends."

Impact on clinical management

Three potential applications of a new learning healthcare system for infectious diseases could address some of the most urgent global problems in infectious disease. One problem is rising antimicrobial resistance, which the World Health Organization names one of the three greatest threats to human health.[51] Care providers overusing antimicrobials and fomenting resistance in subclinical carriers are partly to blame, with recent studies estimating the fraction of misuse to be between quarter to half of all treatments.[52] Multidrug resistance increases the morbidity and mortality of healthcare-acquired infections (HAIs), which have an incidence of 1.7 million cases per year in the US and an estimated annual cost of more than $30 billion[53] that dwarfs the likely cost of any informatics-based preventative efforts. The sobering threat of extensively drug-resistant community-circulating organisms, some of which have therapeutic failure rates of 25-29%,[54] alters the risk analysis for hospital procedures once considered routine and calls for comprehensive new strategies for management.

[51] Infectious Diseases Society of America (2010), "The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020."

[52] McKellar and Fendrick (2014), "Innovation of novel antibiotics: an economic perspective."

[53] Scott (2009), *The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention*.

[54] Hirsch and Tam (2010), "Detection and treatment options for Klebsiella pneumoniae carbapenemases (KPCs): an emerging cause of multidrug-resistant infection."

Identifying high-risk patients for HAI

Infection control for HAIs depends on identifying high-risk patients and applying isolation precautions or reducing known risk factors during their hospital course. For *C. difficile* infection (CDI), the most frequently reported nosocomial infection in the US, many questions about how infections are acquired and managing at-risk patients remain.[55] The prevailing notion that infections are mostly transmitted person-to-person within hospitals[56] conflicts with recent NGS evidence that sources of infection are more diverse,[57] suggesting a greater role for asymptomatic colonized patients and environmental sources.

Each healthcare system represents a unique milieu of person-to-person contact networks, contaminated surfaces, microbiomes, and asymptomatic colonization that contributes to the risk of CDI. EMR and NGS data can prove or disprove transmission between patients and unlock the secrets of modifiable risk factors in this chaotic environment. ML algorithms predicting individual risk of CDI for a large hospital performed better (area under receiver operator curve, AUC=0.81) when operating on >10,000 unconstrained EMR variables rather than curated variables for known risk factors.[58] Similar ML models based on EMR data between 2009-2014 for The Mount Sinai Hospital in New York City, encompassing 192,000 patients and 1,366 CDI diagnoses, show equal performance (AUC=0.80) and draw out associations not typically published for CDI. These may be unique to Mount Sinai's environment and include respiratory failure (odds ratio OR=8.3, 95% confidence interval 6.6–10.3), nutritional irregularity (OR=6.6, 4.7–8.6), and pancytopenia (OR=4.4, 3.1–5.5) (Timothy O'Donnell, personal communication).[59]

A model-based decision support system would screen patients with higher CDI or asymptomatic colonization likelihood and allow earlier diagnosis and intervention. NGS-confirmed transmission events and interactions between people and equipment seen in the EMR and other data could extend this basic model to highlight common factors behind verified transmission and inform empirical, real-time modifications of infection control policy. Cross-sectional analysis by NGS-derived phenotypes and risk factors in the EMR would facilitate more precise clinical decision-making, for instance, whether shortening patient time in intensive care units or decreasing use of provocative antibiotics would be more preventative within the local milieu. Short of a clinical trial that is probably infeasible to conduct, much less replicate across institutions, there is

[55] Leffler and Lamont (2015), "Clostridium difficile."

[56] Cohen et al. (2010), "Clinical practice guidelines for Clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of America (SHEA) and the infectious diseases society of America (IDSA).".

[57] Eyre et al. (2013), "Diverse sources of C. difficile infection identified on whole-genome sequencing."

[58] Wiens, Guttag, and Horvitz (2014).

[59] *These models are developed further in this dissertation to enable propensity modeling for cost analyses in Chapter 5.*

scant evidence for making these decisions at present, so a localized quantitative model can only help.

Earlier detection of outbreaks inside and outside the hospital

Current infection control software suites like VigiLanz Dynamic Monitoring Suite and TheraDoc Infection Control Assistant primarily issue outbreak alerts based on infection frequency thresholds. This could be rendered obsolete by routine NGS of clinical microbiology specimens, which determines with great precision whether a transmission event has occurred.[60] A software system with access to EMR and other hospital data could automatically search elements common between verified transmission cases (caregivers, equipment, or rooms) and alert staff to inspect these elements before they produce enough transmissions to trigger a frequency threshold alert. Given enough historical data, NGS could also help hospitals differentiate community- from hospital-acquired infections and thereby refine metrics used to evaluate infection control policies.[61]

[60] Didelot et al. (2012); Köser et al. (2012).

An active effort to sample the environment inside and outside the hospital could further extend the reach of this surveillance. Within the hospital, "problem spots" identified by earlier investigations could be resampled regularly via NGS to re-evaluate the efficacy of infection control measures. The hospital also samples the pathogen ecosystem of the local population. Hospitals already report diagnoses of highly transmissible and dangerous infections to government authorities, and sharing NGS data for these cases would permit real-time assessment of where pathogens are coming from, how they are evolving, and where populations naïve to a pathogen are located. Current mapping and surveillance efforts[62] would be vastly enhanced by rich phylogenetic information, allowing outbreaks across disparate regions to be linked.[63] Fine-grained, real-time tracking of infectious disease spread would better inform doctors diagnosing and treating new patients, field agents tracking cases and contacts, and health policymakers seeking preventive population measures.

[61] *Software that implements our vision of a genomic clinical microbiology workflow is presented in Chapters 3 and 4 of this dissertation.*

[62] Brownstein et al. (2008), "Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project."

[63] Chin et al. (2011); McAdam et al. (2012), "Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus."; Rasko et al. (2011).

Antimicrobial stewardship

Decision support systems for empirical antibiotic therapy have been investigated for decades,[64] but with the prevalence of antimicrobial resistance skyrocketing, the urgency to implement systems that specifically encourage restraint with antibiotics has increased.[65] Selective reporting is a common strat-

[64] Leibovici et al. (1997), "Improving empirical antibiotic treatment: prospective, nonintervention testing of a decision support system."

[65] Wagner et al. (2014), "Antimicrobial stewardship programs in inpatient hospital settings: a systematic review."

egy that directs providers toward optimal therapies simply by omitting names of inappropriate drugs in susceptibility reports.[66] A more aggressive strategy pushes EMR alerts whenever physicians prescribe antibiotic treatment inconsistent with best practices.[67]

These solutions ignore the power of the EMR to provide evidence that justifies or improves the antimicrobial stewardship interventions. For instance, although it is well accepted that antibiotic overuse increases the prevalence of resistance, current antimicrobial stewardship programs have demonstrated neither effects on patient outcomes nor even that decreased antibiotic treatment leads to decreased antibiotic resistance.[68] By integrating NGS and EMR data, these hypotheses could be investigated in minute detail within large patient cohorts. NGS can reveal and enumerate the genetic mechanisms of resistance circulating through a health system.[69] By tracing the recurrence of pathogens in the local community, an NGS-equipped health system can determine whether patients receiving antibiotics have generated and transmitted drug-resistant mutants. Specific drug regimens can be correlated with the development of particular resistance mutations. Conversely, given enough longitudinal data, the efforts of an antimicrobial stewardship program can be validated by observing decreased emergence of resistance mutations to drugs prescribed more conservatively.

[66] Doern (2013), "Integration of technology into clinical practice."

[67] Kullar et al. (2013), "The "epic" challenge of optimizing antimicrobial stewardship: the role of electronic medical records and technology."

[68] Wagner et al. (2014).

[69] *An example study that uses NGS to find the genetic mechanism for a case of emerging quinolone resistance is presented in Chapter 2 of this dissertation.*

Conclusions

Routine access to pathogen genomic data will transform our ability to manage infections, but only if we can integrate this information with clinical and other data to power predictive models for critical outcomes. Assuming that the hurdles of cost, accuracy, and turnaround time can be addressed, which is likely given current trends, NGS will soon become a standard clinical microbiology procedure. The unprecedented specificity of this data will in the near term allow reconstruction of transmission networks inside and outside of hospitals. In the far term, having rich clinical data linked to pathogen genotypes will permit predictions of prognosis, virulence, and drug susceptibility for active infections once NGS data is available. Incorporating these capabilities into a new clinical workflow that actively refines predictive models by adjusting to new data (Figure 1.2) should improve case management, risk prediction for HAIs, detection

of outbreaks, and antimicrobial stewardship. The missing link in this transformation, and the goal for bringing it to fruition, is software that leverages best-of-breed existing tools, incorporates all relevant heterogeneous datatypes, builds on electronic phenotyping algorithms to scrub low-accuracy EMR data, and validates against gold standard clinical case review. Healthcare institutions and researchers should recognize that a potent combination of NGS and EMR data will transform infectious disease management. The threats posed by multidrug resistance and healthcare associated infections demand a revolution in management strategy. Predictive modeling grounded in rich, diverse molecular and clinical data will dramatically increase the precision of care and help hold these threats at bay.

Notes

An abbreviated version of this chapter was published in *Clinical Infectious Diseases*.[70]

[70] Pak and Kasarskis (2015), "How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management".

Contributions

Theodore R. Pak (TRP) and Andrew Kasarskis (AK) contributed to this chapter. Figure 1.3 is excerpted from Nagar et al. (2014) and Snow (1855). TRP created the remaining figures and wrote the first draft of this chapter. TRP and AK contributed revisions. All contributors saw, revised, and approved the final manuscript. TRP is the first author on the accepted manuscript.

Funding

The authors were supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai.

Conflicts of interest

The authors have no conflicts of interest to disclose.

Acknowledgements

We thank Deena Altman and Shirish Huprikar for critical suggestions on the manuscript.

2

Whole-genome sequencing identifies emergence of a quinolone resistance mutation in a case of Stenotrophomonas maltophilia *bacteremia*

In this chapter we use next-generation sequencing to reveal the mechanism of emerging drug resistance in a case of hospital-acquired infection following the failure of routine antimicrobial therapy. Whole genome sequences for Stenotrophomonas maltophilia *serial isolates from a bacteremic patient before and after development of levofloxacin resistance were assembled* de novo *and differed by one single-nucleotide variant in* smeT*, a repressor for multidrug efflux operon* smeDEF*. Along with sequenced isolates from five contemporaneous cases, they displayed considerable diversity compared against all previously published complete genomes. Whole genome sequencing and complete assembly can conclusively identify resistance mechanisms emerging in* S. maltophilia *strains during clinical therapy.*

You can't win, Darth. If you strike me down, I shall become more powerful than you could possibly imagine.

—OBI-WAN, *Star Wars: Ep. IV – A New Hope*

Life, ah... finds a way.

—IAN MALCOLM, *Jurassic Park*

Stenotrophomonas maltophilia is an aerobic, non-fermenting, and motile Gram-negative bacterium that is increasingly recognized as a cause of hospital-acquired infections with crude mortality rates of 14–69% in cases of bacteremia.[1] Treatment of *S. maltophilia* infections is challenging due to the pathogen's intrinsic resistance to many antibiotic classes via drug efflux pumps, beta-lactamase production, and decreased membrane permeability.[2] Resistance phenotypes are known to change during the course of treatment, which complicates interpretation of automated drug susceptibility testing (DST) results.[3] A mutant strain of *S. maltophilia* with emerging resistance to tetracycline, chloramphenicol, and quinolones was previously characterized following in vitro tetracycline selection.[4] However, little is known about the genetic and molecu-

[1] Brooke (2012), "Stenotrophomonas maltophilia: an emerging global opportunistic pathogen."

[2] Ibid.

[3] Garrison et al. (1996), "Stenotrophomonas maltophilia: Emergence of multidrug-resistant strains during therapy and in an in vitro pharmacodynamic chamber model"

[4] Alonso and Martínez (1997), "Multiple antibiotic resistance in Stenotrophomonas maltophilia."; Sánchez, Alonso, and Martinez (2002), "Cloning and characterization of SmeT, a repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF"

lar mechanisms underlying acquired resistance in the clinical setting—particularly for quinolones, where in contrast to other Gram-negatives, the quinolone-resistance determining region (QRDR) of topoisomerase genes is often unaltered.[5] In this report, we describe the first reported use of whole genome sequencing (WGS) in serial clinical isolates to definitively identify an acquired quinolone resistance mutation in *S. maltophilia*. WGS was performed for the initial and subsequent *S. maltophilia* blood culture isolates from a patient where acquired quinolone resistance was observed (Patient 1) and five other patients (Patients 2-6) from a two-month period in 2013 at The Mount Sinai Hospital.

[5] Valdezate et al. (2005), "Preservation of topoisomerase genetic sequences during in vivo and in vitro development of high-level resistance to ciprofloxacin in isogenic Stenotrophomonas maltophilia strains".

Case report

Patient 1 was a 56 year-old man with a history of pancreatic cancer and a Whipple procedure eleven years earlier who presented to The Mount Sinai Hospital with variceal bleeding at the hepaticojejunostomy site. A transjugular intrahepatic portosystemic shunt was placed, which was complicated by thrombosis. In the following weeks, he had several episodes of polymicrobial bacteremia and was treated with multiple courses of antimicrobials, including a 10-day course of levofloxacin. Two months after levofloxacin exposure, he developed another episode of polymicrobial bacteremia. Blood cultures intermittently grew *S. maltophilia*, *E. faecium*, and *Candida parapsilosis* despite appropriate antimicrobial therapy. Automated DST showed that the first *S. maltophilia* isolate acquired was susceptible to levofloxacin (minimum inhibitory concentration [MIC] 0.5μg/mL) and trimethoprim/sulfamethoxazole (TMP-SMX; MIC ≤20 μg/mL). He was treated with 400 mg intravenous ciprofloxacin every 8 hours, but blood cultures nine days later again grew *S. maltophilia*, now resistant to levofloxacin (MIC >32μg/mL) while still susceptible to TMP-SMX (MIC 1μg/mL). Ciprofloxacin therapy was stopped and intravenous TMP-SMX was given every 8 hours; subsequent cultures did not grow *S. maltophilia*.

Methods

Standard culturing and susceptibility testing for levofloxacin and SXT were performed by automated microbroth dilution with Vitek2® (bioMérieux).

Antimicrobial sensitivities were reported and interpreted according to the 2015 CLSI guidelines for *S. maltophilia*.[6] Isolates were then stocked and frozen at −80°C. Levofloxacin and SXT susceptibilities for all isolates in this study were later confirmed by Etest (bioMérieux) at 24 hours. To prepare for sequencing, isolates were grown from single colonies in tryptic soy broth, and DNA extraction was performed as previously described.[7]

[6] Clinical and Laboratory Standards Institute (2015), *Performance standards for antimicrobial susceptibility testing; twenty-fifth informational supplement M100-S25.*

[7] Altman et al. (2014), "Transmission of Methicillin-Resistant Staphylococcus aureus via Deceased Donor Liver Transplantation Confirmed by Whole Genome Sequencing."

Genome sequencing

Sequencing was performed to a depth of coverage of >150x per genome using the P4-C2 sequencing enzyme and chemistry at the manufacturer's specifications on the PacBio RS II platform (Pacific Biosciences, Menlo Park, CA). For ISMMS2 and ISMMS2R, Sanger sequencing was additionally performed on six PCR-amplified regions encompassing the one single nucleotide variant (SNV) and five one-base indels that differentiated the two PacBio assemblies. Conventional PCR amplification was performed with Choice-Taq Blue (Denville Scientific) and included an initial denaturation step of 180s at 95°C, 30 cycles of denaturation, annealing, and extension at 95°C/30s, 60°C/30s, and 72°C/30s respectively, and a final extension step of 300s at 72°C. Primer sequences are as follows: for the SNV, 5'-`CAAGGTGCTGACCGAAATGC`-3' forward and 5'-`ACACGC CATCCTTCACGTAG`-3' reverse; and for the five indels, 5'-`GCATGGAAGTACCACTGGG T`-3' forward + 5'-`TTGGAGGGGTGGTAAAACGG`-3' reverse, 5'-`TGGCCAACCCCTTCTATG TC`-3' forward + 5'-`CCATGGCCACAGCAAAATGG`-3' reverse, 5'-`CTGCCTTCGGTCACTTC GT`-3' forward + 5'-`TGGAAGTCTCGCTGGAAGGT`-3' reverse, 5'-`GCCCTCTACACCGTCT TTCC`-3' forward + 5'-`GAACTACCGGACGGCTTTGA`-3' reverse, and 5'-`AACTTCTTCGT GTCGGTCCC`-3' forward + 5'-`AGAACTACCGGACGGCTTTG`-3' reverse. Sequences on both strands of the amplified products were determined at an external sequencing facility (Macrogen Inc., Rockville, MD) using the standard Sanger dideoxy-terminator method and the same primers.

Sequence assembly and annotation

Sequencing data was processed and assembled de novo using PacBio's Hierarchical Genome Assembly Process[8] (HGAP, version 3) in the SMRTanalysis toolkit (version 2.3.0) using standard pre-assembly pipeline parameters. Custom scripts were used to circularize the draft assemblies and orient them similarly to reference assemblies K279a, R551-3, D457, and JV3 using the *gyrB* locus

[8] Chin et al. (2013), "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data."

as a landmark.[9] To eliminate overhanging sequence at the end of contigs and to increase accuracy, raw reads were re-mapped to the circularized assemblies using Blasr and the final consensus was re-called using Quiver. Initial annotations were created using the RAST server[10] with specific annotation of sme genes derived from BLAST queries. Depth of coverage reported in Table 2.1 was calculated by SMRTanalysis (version 2.3.0) during re-mapping of reads to the circularized draft assembly.

[9] These scripts are available at `https://github.com/powerpak/pathogendb-pipeline/releases/tag/steno_v1.0` (DOI:`10.5281/zenodo.17295`) within the files `scripts/circularizeContigs.pl` and `scripts/fasta-orient-to-landmark.pl`

[10] Overbeek et al. (2014), "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).".

Accession numbers

Sequences and annotations for reference assemblies of clinical *S. maltophilia* isolates K279a, R551-3, D457, and JV3 were obtained from GenBank/RefSeq at accession numbers `AM743169.1`, `NC_011071.1`, `NC_017671.1`, and `NC_015947.1`, respectively. These represent the entirety of assemblies for *S. maltophilia* found in NCBI Assembly with an Assembly Level of "Complete Genome" (`http://www.ncbi.nlm.nih.gov/assembly/organism/40324/all/`) at the time of the study.[11] K279a and D457 were isolated from human infections, while R551-3 and JV3 were isolated from plants. Previously published sequences for the quinolone-resistance determining region (QRDR) of the *gyrA*, *gyrB*, *parC* and *parE* genes in *S. maltophilia*[12] were obtained from EMBL/European Nucleotide Archive.

[11] By 2017, excepting the three assemblies submitted as a result of this study, only one more complete genome (ASM202560v1) is available.

[12] Valdezate et al. (2002), "Topoisomerase II and IV quinolone resistance determining regions in Stenotrophomonas maltophilia clinical isolates with different levels of quinolone susceptibility".

Complete genome sequences for ISMMS2, ISMMS2R, and ISMMS3 were deposited in GenBank at accession numbers `CP011305`, `CP011306`, and `CP011010`, respectively. Deposited sequences for ISMMS2 and ISMMS2R incorporate the Sanger corrected regions described above. Sequences for ISMMS4, ISMMS5, ISMMS6, and ISMMS7 were deposited as Whole Genome Shotgun projects at DDBJ/EMBL/GenBank under the accessions `JZIU00000000`, `JZIV00000000`, `JZIW00000000`, and `JZTX00000000`, respectively, with the versions in this chapter at `JZIU01000000`, `JZIV01000000`, `JZIW01000000`, and `JZTX01000000`, respectively.

Comparative genomic analysis

Pairwise comparison between strains was performed with the MUMmer 3.23 package,[13] firstly using nucmer for pairwise genome alignment. The resulting nucmer alignments were filtered for quality and uniqueness via the `delta-filter` tool (using the `-1` flag to identify top alignments between the refer-

[13] Delcher, Salzberg, and Phillippy (2003), "Using MUMmer to identify similar regions in large sequence sets."

ence and query intervals). To estimate phylogenetic tree distances, high-quality SNP and indel calls were assigned via the `show-SNPs` tool using the `–C` flag to only report SNPs in regions with unambiguous mappings. For ISMMS2 and ISMMS2R, `show-SNPs` was also used without the `–C` flag to verify that no additional SNPs or indels were in ambiguously mapped regions.

Mugsy version 2.2 was used to perform multiple sequence alignment of the whole genome sequences in order to find local collinear blocks (LCBs) of conserved sequence.[14] These aligned blocks were used to establish a core genome (of 3.01 Mbp) across all isolates, from which a phylogenetic tree was constructed using RAxML version 8.0.2,[15] employing the GTRGAMMA substitution model and performing 20 runs. Whole genome alignments for visualization of recombination events was performed with Mauve 2.4.0,[16] using the *progressiveMauve* algorithm[17] with a minimum seed weight of 21, seed families enabled, and all other parameters at defaults. Clustal Omega[18] was used for multiple sequence alignment of putative amino acid sequences, which were then rendered with ESPript version 3.0.[19]

[14] Angiuoli et al. (2011), "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing."

[15] Stamatakis (2014), "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies."

[16] Darling et al. (2004), "Mauve: multiple alignment of conserved genomic sequence with rearrangements."

[17] Darling, Mau, and Perna (2010), "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement".

[18] Sievers et al. (2011), "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega".

[19] Robert and Gouet (2014), "Deciphering key features in protein structures with the new ENDscript server".

Epigenetic motif analysis

For each isolate, initial DNA modification motifs were first predicted by a de novo motif discovery pipeline in SMRTportal (`RS_Modifications_Motif_Analysis.1`). The pipeline searches for kinetic variations in DNA polymerization events recorded during sequencing that correlate with modifications in the template, with different modifications creating distinct kinetic profiles.[20] At the coverage depths reported for this study, the probability (power) of detecting a modification event at a site at the 0.1 significance threshold, if it is truly modified, exceeds 99.99%.[21] Raw predictions, which often have incorrectly- or over-called motifs, were further refined by a re-analysis of the raw data using a single molecule level characterization method.[22] Conceptually, this method was used to check the single molecule level methylation status of each putative motif and its neighboring (more or less specific) motifs and determine the real motif.

[20] Clark et al. (2012), "Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing"; Fang et al. (2012), "Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing".

[21] Fang et al. (2012).

[22] Beaulaurier et al. (2015), "Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes".

Results

Two complete whole genome sequences were derived from Patient 1's isolates before and after the change in levofloxacin MIC and compared to whole genome sequences of five control *S. maltophilia* isolates (Patients 2-6). All sequences were *de novo* assembled, i.e., without regard to reference assemblies. Table 2.1 summarizes the relative dates of collection, antimicrobial susceptibility results, and assembly statistics.

Patient	Time of collection (days)[a]	Isolate name	Levo susceptibility (MIC, mg/L)		SXT susceptibility (MIC, mg/L)		Assembly quality	Depth of coverage
			Vitek2	Etest	Vitek2	Etest		
1	0	ISMMS2	S (0.5)[b]	S (1)	S (<20)	S (0.19)	1 circular 4.51Mbp chromosome	160×
1	+10	ISMMS2R	R (>32)[b]	R (16)	S (1)	S (0.38)	1 circular 4.51Mbp chromosome	403×
2	−26	ISMMS3	S (0.25)	S (0.38)	U (80, <20)[c]	S (0.75)	1 circular 4.80Mbp chromosome	153×
3	+14	ISMMS4	R (>8)	R (>12)	U (0.5, 80)[c]	S (0.75)	3 contigs (4.73Mbp, 6.5kbp, 11.2kbp)	303×
4	−32	ISMMS5	S (1)	S (1)	S (<20)	S (0.25)	18 contigs	270×
5	0	ISMMS6	S (<0.12)	S (0.125)	S (<20)	S (1.5)	10 contigs	262×
6	+2	ISMMS7	S (1)	S (0.75)	S (<20)	S (1.5)	1 circular 4.69Mbp chromosome, 1 additional 17.7kbp contig	318×

Table 2.1: **Sequenced clinical isolates and their antimicrobial susceptibilities.** Abbreviations: Levo, levofloxacin; SXT, trimethoprim/sulfamethoxazole; S, susceptible; R, resistant; U, undetermined; Mbp, million base pairs; kbp, thousand base pairs.

[a] Time of collection was defined in days relative to the date of collecting the initial *S. maltophilia* isolate in the case patient

[b] This is the change in levofloxacin susceptibility investigated in this study.

[c] Inconsistent results were obtained in replicate.

Emergence of a point mutation conferring quinolone resistance

Assembled genome sequences for Patient 1's isolates before (ISMMS2) and after (ISMMS2R) observation of levofloxacin resistance were compared directly and were identical except for one single-nucleotide variant (SNV) and five one-base indels. Sanger sequencing confirmed the presence of the SNV, but identified the indels as homopolymer assembly errors. Coding domain sequence predictions for the surrounding locus (Figure 2.1A) revealed that the SNV was inside *smeT*, a *tetR*-like repressor upstream of the structural operon for the *smeDEF* genes, which encode a multidrug efflux pump. The SNV is an A>T substitution at position 497 of *smeT* causing a nonsynonymous Leu-166→Gln mutation.

The same nonsynonymous mutation has been previously observed in an in

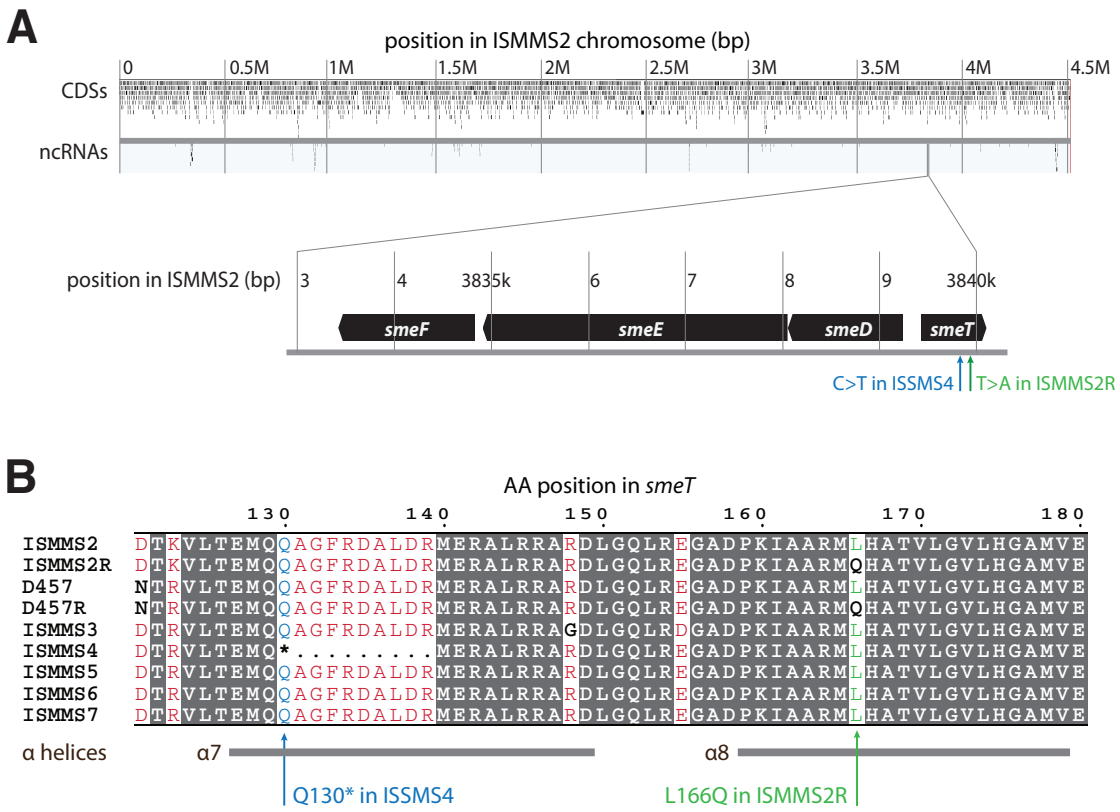


Figure 2.1: **Single-nucleotide variants (SNVs) observed in quinolone-resistant *S. maltophilia* clinical isolates.** A, assembled circular chromosome for ISMMS2, including predicted coding domain sequence (CDS) and noncoding RNA (ncRNA) features drawn with ChromoZoom. Horizontal position corresponds to base pair location. The *smeDEF* operon is shown in the detail callout, which highlights both the *smeT* c.497T>A SNV that emerged in ISMMS2R and the aligned location of the *smeT* c.388C>T SNV (encoding a premature stop codon) in ISMMS4. ISMMS2 and ISMMS2R are serial isolates from a single patient before and after development of quinolone resistance, while ISMMS4 was quinolone-resistant at initial isolation from a different patient. B, multiple sequence alignment of part of the predicted *smeT* product in each of the clinical isolates, the D457 reference assembly, and its quinolone resistant counterpart D457R. Predicted α-helices are labeled as grey bars below the sequence. Positions identical in all sequences are shaded with a dark gray background, equivalent substitutions are typeset in red, and non-equivalent substitutions are typeset in boldface black. The L166Q and Q130* (*, stop codon) polymorphisms are highlighted.

vitro strain of *S. maltophilia*, D457R, created by selecting single-step tetracycline-resistant mutants from the antibiotic-susceptible clinical strain D457.[23] The mutation is in the eighth α-helix of the *smeT* protein,[24] which homodimerizes to repress transcription of the *smeDEF* operon.[25] Although the mutation is not in the DNA-binding region, it has been shown to disable the repressor activity of *SmeT*,[26] leading to upregulation of *SmeDEF* and conferring an MDR phenotype.[27]

Figure 2.1B shows an amino-acid sequence alignment comparing *SmeT* in D457 and D457R to aligned sequences from our seven isolates. Notably, while none of the remaining isolates shared the same Leu-166→Gln (c.497A>T) mu-

[23] Alonso and Martínez (1997); Sánchez, Alonso, and Martinez (2002).

[24] Hernández et al. (2009), "Structural and functional analysis of SmeT, the repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF".

[25] Hernández et al. (2009); Sánchez, Alonso, and Martinez (2002).

[26] Sánchez, Alonso, and Martinez (2002).

[27] Alonso and Martínez (2001), "Expression of multidrug efflux pump smeDEF by clinical isolates of Stenotrophomonas maltophilia".

tation, another isolate resistant to levofloxacin, ISMMS4, displayed a C>T mutation at position 388 of *smeT* that creates a premature stop codon that likely disrupts *smeT* function (Figure 2.1A and 2.1B).

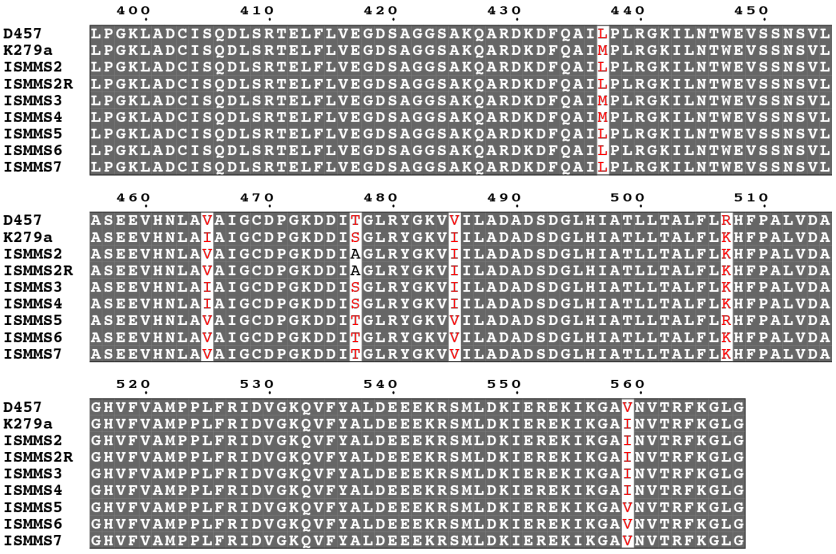


Figure 2.2: Amino-acid sequence alignment for the quinolone-resistance determining region (QRDR) of the *parE* gene for seven S. maltophilia clinical isolates (ISMMS2 through 7 and ISMMS2R) and two reference assemblies of clinical isolates obtained from GenBank.

The QRDR are loci within genes encoding topoisomerase II and IV subunits known for mutations that confer quinolone resistance in Gram-negative bacteria, although they appear to play a secondary role to efflux systems for resistance emerging during treatment of *S. maltophilia* infection.[28] An amino-acid sequence alignment of the *gyrA*, *gyrB*, and *parC* genes of our seven isolates and the reference clinical isolates D457 and K279a revealed no differences in the QRDR. Some variants were observed within the QRDR of *parE* (Figure 2.2), all of which were consistent with past observations in clinical isolates[29] except for an Ile-599→Val variant observed in three of our isolates and the D457 reference sequence.

[28] Valdezate et al. (2005).

[29] Valdezate et al. (2002).

Diverse sources of S. maltophilia *identified with WGS*

Significant genomic diversity was observed among the *S. maltophilia* isolates from all six patients. Figure 2.3 shows a maximum-likelihood phylogeny with branch lengths scaled to SNV distances. Our isolates distribute widely among all four reference assemblies for complete *S. maltophilia* genomes in GenBank. The distances of tens of thousands of SNVs seen in our phylogeny

suggest that the natural diversity of pathogenic *S. maltophilia* is greater than that captured by the current set of reference assemblies, even within a single hospital setting.

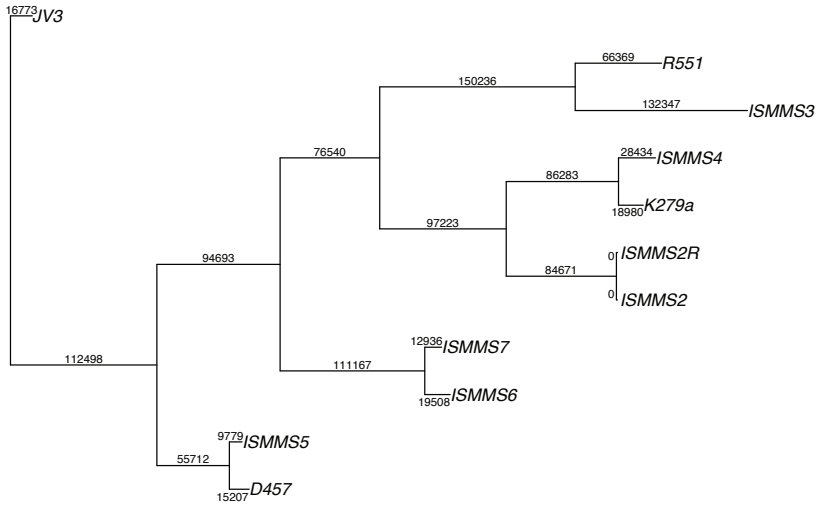


Figure 2.3: **Phylogeny of seven *S. maltophilia* clinical isolates (ISMMS2 through 7 and ISMMS2R) and four reference assemblies obtained from GenBank.** Trees were constructed by inferring ancestral states using RAxML-8.0.2; branch lengths correspond to single-nucleotide polymorphism (SNP) distances from branch points, and are drawn using R version 3.0.3 and the APE library version 3.1-1. The core genome did not contain the *smeT* locus; therefore, the SNV differentiating ISSMS2 and ISMMSR is not observed in this tree.

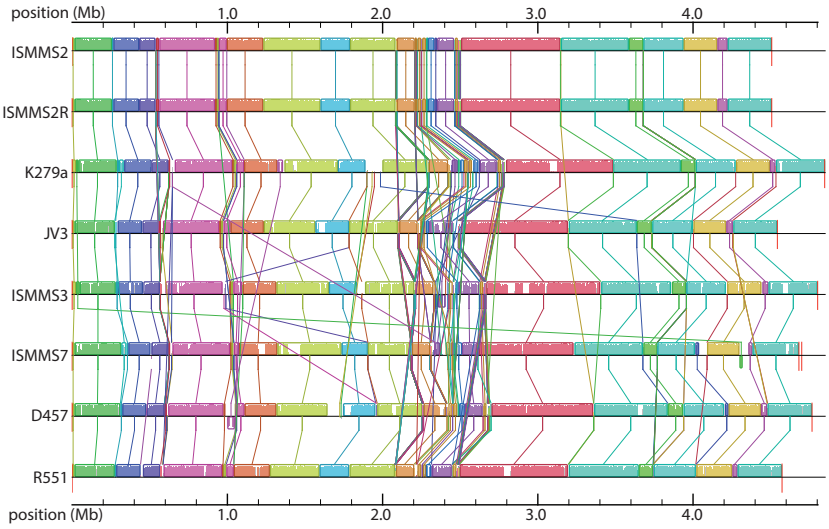


Figure 2.4: **Genome-scale comparison of four fully assembled *S. maltophilia* clinical isolates and four reference assemblies obtained from GenBank.** Mauve 2.4.0 was used to plot locally collinear blocks (LCBs; conserved segments that appear to be internally free from genome rearrangements) as colored rectangles, with gaps representing non-homologous regions. Vertical bars inside each LCB rectangle show the average level of conservation at that region of the genomic sequence. Colored lines connect homologous LCBs among the genomes, and LCBs plotted below the centerline are in the reverse complement orientation relative to the ISMMS2 sequence. At top, sequences for the isolates from before and after development of quinolone resistance (ISSMS2 and ISSMS2R) in the case patient have identical structures.

Recombination is not an obvious source of diversity in our *S. maltophilia* isolates. Figure 2.4 depicts whole genome alignments between the four clinical isolates where assembly produced a circularized chromosome and the four GenBank references, showing small areas of non-homology separating large regions of significant homology occurring generally in the same order for each

genome. ISMMS2 and ISMMS2R are structurally identical, as expected for serial isolates, while recombination events among other strains are limited to small 1-2kb segments. Epigenetics motif analysis also suggests that the isolates are not related. Table 2.2 shows different motifs in isolates from separate patients, implicating differences in type II & III restriction modification systems between the isolates more likely to be caused by inter-strain/species horizontal transfer of methyltransferases than by intra-strain mutations.[30] Together, this demonstrates that transmission did not occur among these six cases and that whole-genome sequencing can comprehensively capture genetic distances and structural variants among diverse clinical isolates of *S. maltophilia*.

[30] Srikhanta, Fox, and Jennings (2010), "The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes."

Isolate name	Epigenetic motifs
ISMMS2	AGTACT
ISMMS2R	AGTACT
ISMMS3	None
ISMMS4	CAGAG
ISMMS5	CTGGAC, CACANAG
ISMMS6	CAACAC, CTGATG, CAACGAC
ISMMS7	CAGAG

Table 2.2: **Diverse epigenetic motifs**, representing putative target sequences for each strain's DNA methyltransferase enzymes, discovered for clinical isolates of *S. maltophilia*. Isolates are named as in Table 2.1. The underlined A's correspond to putative 6-methyladenine residues, which was the only modification type found in this study.

Discussion

This is the first report of WGS on serial isolates to characterize the emergence of a resistance mutation in *S. maltophilia* during antibiotic treatment of an active infection. In contrast to studies sequencing highly resistant strains of *S. maltophilia* to reveal various intrinsic and acquired antibiotic resistance genes,[31] where it remains difficult to assess their relative importance to the phenotype, performing WGS on serial isolates as resistance emerges in vivo allows the causative mutation(s) to be captured. In our patient, the mutation was a SNV that replicates a variant observed in an in vitro model strain created to study the MDR phenotype in 1997.[32] Using WGS and susceptibility testing, we can confirm that this SNV was the only variant to emerge and that it was sufficient to confer quinolone resistance in a clinical case. This underscores the need for clinicians to consider repeating DST during monotherapy if clinical signs suggest therapy failure.

[31] Crossman et al. (2008), "The complete genome, comparative and functional analysis of Stenotrophomonas maltophilia reveals an organism heavily shielded by drug resistance determinants."; Zhao et al. (2015), "Identification and characterization of a serious multidrug resistant Stenotrophomonas maltophilia strain in China."

[32] Alonso and Martínez (1997).

smeT appears to play a central role in adaptive resistance to quinolones and

other antibiotics effluxed by *smeDEF*, like tetracycline, chloramphenicol, erythromycin, and aminoglycosides. Since any mutation that inactivates this protein would be able to derepress *smeDEF* and confer resistance, *smeT* is under intense selective pressure in the presence of these drugs. In this study, we observed not only a deleterious SNV in the strain that displayed resistance (ISMMS2R), but a premature stop codon in *smeT* in a strain that was already resistant at first isolation (ISMMS4). Certain nucleotide positions appear to be under greater selective pressure than others, as evidenced by our observation of the same mutation that occurred in D457R, and a relative paucity of nonsynonymous coding mutations in *smeT* observed among clinical *smeT* isolates.[33] Since sustained overexpression of *smeDEF* is physiologically unfavorable,[34] it is possible that pathogenic strains of *S. maltophilia* rely on natural diversity of mutations in the *smeT* locus to activate or deactivate *smeDEF* expression, allowing for rapid adaptation to antibiotic stress, though further study is needed.

Since resistance from a single SNV emerged during a short course of ciprofloxacin, clinicians should be cautioned about using quinolone monotherapy for *S. maltophilia* bacteremia, as highlighted in recent retrospective studies.[35] The wide variety of MDR phenotypes and unreliability of DST results has created uncertainty about appropriate treatment for *S. maltophilia*, but SXT remains the most common choice for monotherapy.[36] SXT resistance in *S. maltophilia* is not known to be caused by efflux systems but has been linked to Class 1 integrons and ISCR elements.[37] This suggests that spontaneous resistance is less likely to emerge with SXT monotherapy, although a clinical trial comparing the two antibiotics is warranted.[38]

In conclusion, characterizing the full extent of genetic alterations that *S. maltophilia* utilizes to develop antibiotic resistance in vivo and improving genomic surveillance of clinical strains will help refine antibiotic selection criteria available to clinicians. This study furthermore highlights the utility of WGS for profiling the precise mutations underlying emerging antibiotic resistance in clinical cases of bacteremia. Although short read WGS has previously revealed evolving mechanisms of resistance within infections by relatively well-studied species like *Staphylococcus aureus*[39] and *Mycobacterium tuberculosis*,[40] here, we have shown that long read WGS and *de novo* assembly permit the investigation of species with few available reference genomes.

[33] Sanchez, Alonso, and Martinez (2004), "Regulatory regions of smeDEF in Stenotrophomonas maltophilia strains expressing different amounts of the multidrug efflux pump SmeDEF".

[34] Alonso et al. (2004), "Overexpression of the multidrug efflux pump SmeDEF impairs Stenotrophomonas maltophilia physiology".

[35] Cho et al. (2014), "Can levofloxacin be a useful alternative to trimethoprim-sulfamethoxazole for treating Stenotrophomonas maltophilia bacteremia?"; Wang et al. (2014), "Monotherapy with fluoroquinolone or trimethoprim-sulfamethoxazole for treatment of Stenotrophomonas maltophilia infections".

[36] Brooke (2012); Cho et al. (2014); Wang et al. (2014).

[37] Brooke (2012).

[38] Cho et al. (2014); Wang et al. (2014).

[39] Mwangi et al. (2007), "Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing."

[40] Comas et al. (2011), "Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes".

Notes

An abbreviated version of this chapter was published in *Antimicrobial Agents and Chemotherapy*.[41]

[41] Pak et al. (2015), "Whole-Genome Sequencing Identifies Emergence of a Quinolone Resistance Mutation in a Case of Stenotrophomonas maltophilia Bacteremia."

Contributions

Theodore R. Pak (TRP), Deena R. Altman (DRA), Oliver Attie (OA), Robert Sebra (RS), Camille L. Hamula (CLH), Martha Lewis (ML), Gintaras Deikus, (GD) Leah C. Newman (LCN), Gang Fang (GF), Jonathan Hand (JH), Gopi Patel (GP), Fran Wallach (FW), Eric E. Schadt (EES), Shirish Huprikar (SH), Harm van Bakel (HVB), Andrew Kasarskis (AK), and Ali Bashir (AB) contributed to this chapter.

TRP, DRA, SH, and AK designed the study. DRA, GP, FW, and CLH organized collection of the samples. GD, LCN, and RS prepared sequencing libraries and performed sequencing. CLH performed culturing and drug susceptibility testing. ML performed PCR for Sanger sequencing. TRP, OA, RS, GF, HVB, and AB performed data analysis. EES, SH, FW, and AK provided institutional support. JH wrote the first draft of the case report. TRP created all figures and wrote the first draft of this chapter. All contributors saw, revised, and approved the final manuscript. TRP is the first author on the accepted manuscript.

Funding

Funding was provided by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai, and also in part by the NIAID-supported NRSA Institutional Research Training Grant (5T32-AI764713) for Global Health Research (DRA). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

We thank Timothy O'Donnell, Tavi Nathanson, Jose Clemente, Flora Samaroo, Angelo Rendo, and members of the clinical microbiology laboratory at Mount

Sinai for their contributions. This work was supported in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

3

ChromoZoom v2: dynamic, online visualization of genomes and next-generation sequencing data

Although new sequencing technologies enable the automatic assembly of complete genomes from long reads and the generation of copious layers of functional data, sharing and collaboratively exploring these datasets remains difficult with current software. Genome browsers are the standard approach for interactively visualizing these data and may also facilitate collaboration if they are accessible via the web. The first version of ChromoZoom was written in 2012 as a fast, fluid web interface for genome browsing, although the design centered around scraped data for three reference genomes. With the present pace of de novo *assembly and resultant growth in the number of references, this strategy requires some rethinking. Here, I present a re-engineered version of ChromoZoom that can dynamically load custom genome assemblies and related data and also adds many new features that support pathogen surveillance activities at Mount Sinai.*

Unfortunately, no one can be told what the Matrix is. You have to see it for yourself.

—MORPHEUS, *The Matrix*

OFFICER RAMATHORN: [Using a computer] *Enhance. Enhance. Enhance.*

CAPTAIN O'HAGAN: *Just print the damn thing!*

—*Super Troopers*

GENOME BROWSERS ARE INDISPENSIBLE tools for modern experimental and computational biology. The size of typical genomic data (anywhere from tens of thousands of nucleotides for small viruses into the billions for humans) prevents static plots from simultaneously capturing their scale and detail in a reasonable amount of space. Therefore, genome browsers are tasked with laying out the data in a way that is both interpretable and easily navigable. Most designs use a movable viewing region and map the contigs (in a completed assembly, the largest of which are the organism's chromosomes) to one axis of the coordinate system, usually a horizontal axis, while stacking the datasets onto the other axis. These other data can include range features such as genes and coding regions, continuous quantitative data like read depth and methylation

levels, and mappings of other sequences, such as a different genome or the reads from a sequencing experiment. Many other datatypes exist and continue to be invented, but these are the basic shapes of data that genome browsers are now expected to handle.

Current genome browsers fall into three major categories, each with their own advantages and disadvantages. In fetching and drawing large amounts of heterogenous data, *desktop-based browsers* like IGV,[1] IGB,[2] and SmrtView[3] have retained an advantage in being fastest during typical navigation operations, particularly when the data is being read from the user's local hard disk. All of the aforementioned desktop applications, however, use Java; this incurs the penalty of installing a Java VM in addition to the browser itself, and precludes their use on most mobile devices. Because of the need to download software, these applications are not well-equipped for the rapid sharing of an active browsing session with a colleague, as website users are quite accustomed to doing by sending a URL or using "share" buttons. Nevertheless, they remain very popular for browsing large alignments from typical next-generation sequencing (NGS) experiments on humans and model organisms, since the size (and potential confidentiality) of these data often prevent them from being sent over the internet into web-based genome browsers.

[1] Thorvaldsdóttir, Robinson, and Mesirov (2013), "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration".

[2] Freese, Norris, and Loraine (2016), "Integrated genome browser: Visual analytics platform for genomics".

[3] `https://github.com/PacificBiosciences/DevNet/wiki/SMRT-View`

The first generation of *web-based* genome browsers like UCSC,[4] GBrowse,[5] Ensembl,[6] and the NCBI Map Viewer[7] continue to retain the advantage of instant access from any web browser and the simplicity of sharing a particular view of the genome by linking to it. Since they were some of the first genome browsers to gain widespread use by wet lab biologists, and offered centralized repositories of valuable data during key moments of the birth of human genomics, they still tend to exhibit the widest array of expert-curated data. UCSC's browser is well known for being the first to provide rapid, interactive access to early products of the Human Genome Project,[8] and remains prominent in human genomics for providing thousands of widely-used tracks that cover everything from tissue-specific expression and evolutionary conservation to clinically significant variants. However, since these software projects were built before modern web technologies like Asynchronous Javascript and XML (AJAX)[9] and HTML5, the server must draw each user's data as an image that is sent to their browser, and the software's interactivity is therefore constrained compared to newer web applications. Unlike, e.g., Google Maps, none of these

[4] Dreszer et al. (2011), "The UCSC Genome Browser database: extensions and updates 2011".

[5] Stein (2002), "The Generic Genome Browser: A Building Block for a Model Organism System Database".

[6] Stalker et al. (2004), "The Ensembl Web site: mechanics of a genome browser."

[7] Wheeler et al. (2003), "Database resources of the national center for biotechnology".

[8] Kent et al. (2002), "The Human Genome Browser at UCSC".

[9] Paulson (2005), "Web Applications with Ajax".

browsers allow the user to smoothly scroll, zoom, or "throw" the viewport (also called inertial scrolling), and they cannot draw new data without interrupting the user's movements. This latency makes it frustrating to gain an intuitive feel for distances or zoom levels and clashes with the experience users expect from the way scrolling[10] works on most other apps and websites.

A new generation of web-based genome browsers emerged around 2009 that embraced AJAX and employed JavaScript to draw data on the client-side. Of these, projects that have stayed active include JBrowse,[11] Dalliance,[12] and Anno-J.[13] Among their advantages were a level of interactivity that approached the look and feel of the desktop-based browsers, without a need to download and install any software. Additionally, in some cases, views of the data could be shared as in the older web-based browsers. However, the new browsers were largely divorced from the core service of track curation that was integral to older web-based genome browsers, providing only barebones demo sites with a small number of sample tracks. Including the most recent entrants in this category, pileup.js[14] and igv.js,[15] all of the next-generation genome browsers are provided as source code libraries, which expect the user to install the code to their own web server and marry it to data via configuration or embedding components into a larger website. Obviously, this reduces the accessibility of the software to teams with a web developer and a webserver.

There still is no genome browser that hits all the high marks of Google Maps in providing a (1) universally accessible, (2) immersively interactive, and (3) easily sharable visualization of large, multilayered datasets. Without requiring any additional software, Google Maps even allows users to add custom data that is plotted on top of a map, and then share it via URL or embed it into another site.[16] Although Dalliance and JBrowse do offer some user interfaces (UIs) for layering and configuring user-provided tracks alongside server-provided data, they don't offer any sharing capabilities for these "mashup" views. In fact, none of the newer generation of web-based genome browsers emphasize this use case, assuming the involvement of a server administrator who can configure custom data sources. Finally, despite the increasing production of new genome layouts (or reference assemblies) as *de novo* assembly of NGS data becomes commonplace, none of the new generation of web-based genome browsers allow a user to dynamically load a custom assembly (with annotation and sequence data) via their UI.[17] Therefore, there are still avenues for improving web-based ge-

[10] And pinching, since the release of the iPhone. Both operations on modern smartphones take pains to stay at a buttery smooth 60 frames per second, even if rendering quality has to be temporarily downgraded.

[11] Buels et al. (2016), "JBrowse: a dynamic web platform for genome visualization and analysis".

[12] Down, Piipari, and Hubbard (2011), "Dalliance: interactive genome viewing on the web."

[13] http://www.annoj.org

[14] Vanderkam et al. (2016), "pileup.js: a JavaScript library for interactive and in-browser visualization of genomic data".

[15] https://github.com/igvteam/igv.js

[16] An example of a running route in New York City: https://goo.gl/GoHxx8

[17] Even Google Maps allows loading of a different GIS "reference," i.e., a non-Earth planetary coordinate system: see https://www.google.com/moon/ and https://www.google.com/mars/

nome browsers to reach the level of the user experience that Google Maps first demonstrated possible in 2004 for geospatial data.[18]

[18] Vincent (2007), "Taking online maps down to street level".

Implementation

The first version of ChromoZoom,[19] released in 2012, attempted to achieve the three previously enumerated design goals, but targeted a small set of reference genomes in UCSC's database: GRCh37/hg19 (Human, Feb. 2009), NCBI36/hg18 (Human, Mar. 2006), and sacCer3 (Baker's yeast, Apr. 2011). At the time, I pursued a strategy of scraping UCSC's rendered images for these genome tracks at multiple zoom levels and serving them as tiles, similar to the approach of the first version of Google Maps.[20] Our approach ran into several roadblocks over time. Firstly, due to space constraints, it was impossible to scrape tiles for all of the thousands of tracks that UCSC hosts, so I had to focus only a small, frequently used subset for the two human assemblies. Secondly, since it was so computationally expensive to render all the tiles for a new genome, there was no expectation that the user could ever load a previously unseen genome assembly (or request that one be loaded) from the client side. Finally, the cost of re-rendering tiles made it difficult for us to keep tracks up to date with UCSC's data, even after I created a local mirror of the UCSC genome browser. Although ChromoZoom did allow for user-provided custom data to be drawn alongside the tiled images using client-side JavaScript and `<canvas>` elements, I expected that server-hosted tracks would remain most popular.

[19] Pak and Roth (2013), available at `http://chromozoom.org/`

[20] See Skinner et al. (2009). Today, Google Maps renders all data using WebGL polygons when available, which allows for graphics processing unit (GPU) acceleration in the browser and high framerates.

As demand for viewing different genome assemblies and more varieties of custom data grew, I realized that our server-side tile scraping strategy would be untenable in the long run. Particularly in the context of pathogen surveillance and microbiology, where the number of completed reference genomes for certain species is already in the hundreds,[21] I discovered a general need for a web-based genome browser that could dynamically load a custom genome assembly with associated sequence, alignment, and annotation files. Although desktop-based genome browsers (particularly IGB) are able to open custom assemblies, their visualizations are not easily shared amongst a diverse team of wet lab biologists, bioinformaticians, and clinicians.

[21] As of April 2017, there are 154 complete *Staphyloccocus aureus* genomes on NCBI Genome: `https://www.ncbi.nlm.nih.gov/genome/genomes/154`

Therefore, I re-engineered ChromoZoom in its second release (v2) to

support a new strategy of displaying every genome entirely via the use of `<canvas>` and scalable vector graphics (SVG), with client-side parsing and display of all contig layout, track annotation, and sequence data from standard file formats. The efficient fetching of track data that falls within the user's current viewport is facilitated by the use of bigBed and bigWig formats (big* formats),[22] which use binary compression that exceeds the storage efficiency of our previous tiled images (in PNG format). To test whether this new strategy scales, I scraped all active genomes from UCSC using a custom pipeline written in Python, which converts data from UCSC's MySQL database into bigBed formatted files or links directly to files hosted at UCSC that can be displayed directly by ChromoZoom. Henceforth I provide all UCSC genome data via this mechanism and not as tiled images.

[22] Kent et al. (2010), "BigWig and BigBed: enabling browsing of large distributed datasets".

ChromoZoom is still constructed as a single-page web application that implements most features in JavaScript, with heavy use of the jQuery UI library, and compilation into a single minified bundle using `browserify`.[23] All genomic data is fetched via AJAX from PHP scripts on the server that interface with libraries and utilities for each genomic data format, sometimes retrieving the backing data from another server.

[23] `browserify` is a node.js package that can combine modules in separate JavaScript source files for use by web browsers. See `http://browserify.org/`

For client-side performance, I continue to use tiled HTML5 elements that are rescaled and moved in accordance with user pan, zoom, and throw operations. I prerender into off-screen tiles to reduce loading latency and maintain a sense of location even when the viewport moves or zooms. For maximal performance when animating elements during viewport movement, I use the GreenSock Animation Platform.[24] To the greatest extent possible, fetching and parsing of annotation tracks and genome data occurs within Web Workers, which offload execution of JavaScript into processes that don't block the browser's main UI thread; this is a unique performance advantage of ChromoZoom over the other next-gen genome browsers.[25] Although most rendering is done in rasterized `<canvas>` tiles rather than vector graphics for performance reasons, I fully support high resolution (a.k.a. high-DPI or "retina") displays by always rendering at the native pixel resolution of the display device. All figures in this chapter are screenshots from a high-DPI display.

[24] GreenSock uses CSS3 transforms to enable hardware acceleration for most animations. See `https://greensock.com/`

[25] Buels et al. (2016); Down, Piipari, and Hubbard (2011); Vanderkam et al. (2016).

Availability

ChromoZoom's source code is available from GitHub at https://github.com/rothlab/chromozoom. ChromoZoom v2 can be viewed online by nearly any modern web browser (Chrome, Safari, Firefox, or IE11+) at https://pakt01.u.hpc.mssm.edu/chromozoom/.[26]

All source code is available under an AGPLv3 license[27] and is free for academic and personal (but not commercial) use. For those that want to develop on the ChromoZoom codebase or deploy it locally (e.g., to view or serve data within a firewall), it can be installed to any Linux or macOS server with Apache web server, PHP 5.x, and a set of executables from `htslib` and Jim Kent's utilities, both of which are freely downloadable.[28] Optionally, the user can also compile and install `bigBedSearch`, which is a separate software project created by the author[29] that adds better searching capabilities for bigBed formatted data.

[26] This will be moved to http://chromozoom.org/, replacing v1, for the public release.

[27] Briefly, this means that you can use ChromoZoom code for derivative applications, as long those applications are also open source, even if they are provided as web services. See https://www.gnu.org/licenses/agpl-3.0.en.html

[28] See http://www.htslib.org/download/ and http://hgdownload.cse.ucsc.edu/admin/exe/.

[29] See https://github.com/powerpak/bigBedSearch

Results and Discussion

Data mirrored from UCSC

183 genomes from UCSC were scraped for all track types compatible with ChromoZoom v2, which included bigWig, bigBed, BAM,[30] tabix-compressed VCF,[31] wiggle, and BED, and all track types readily convertible to the bigBed format, which included genePred, rmsk, PSL, GVF, and narrowPeak.[32] This produced 8,702 converted bigBed files and 7,670 additional track entries that point directly to big* and BAM files on UCSC's servers. Collectively, the scraped tracks consume 485 GB of disk space. Since UCSC's MySQL database contains information on the update time for each table, I can re-run this process on a weekly schedule and only re-download data for tracks that have been updated, which occurs infrequently for almost all tracks. I therefore can efficiently keep the local bigBed files continuously synchronized with UCSC's database.

[30] Li et al. (2009), "The Sequence Alignment/Map format and SAMtools".

[31] Danecek et al. (2011), "The variant call format and VCFtools"; Li (2011), "Tabix: fast retrieval of sequence features from generic TAB-delimited files".

[32] Descriptions of all formats used by UCSC can be found at https://genome.ucsc.edu/FAQ/FAQformat.html

Browsing a custom bacterial genome assembly

The browser interface of ChromoZoom v2 is shown in Figure 3.1 visualizing an assembled genome for an isolate of *Staphyloccocus aureus* from a patient at The Mount Sinai Hospital.[33] The interface uses the top line to display

[33] The isolate was sequenced and assembled using the methods described in Chapter 2.

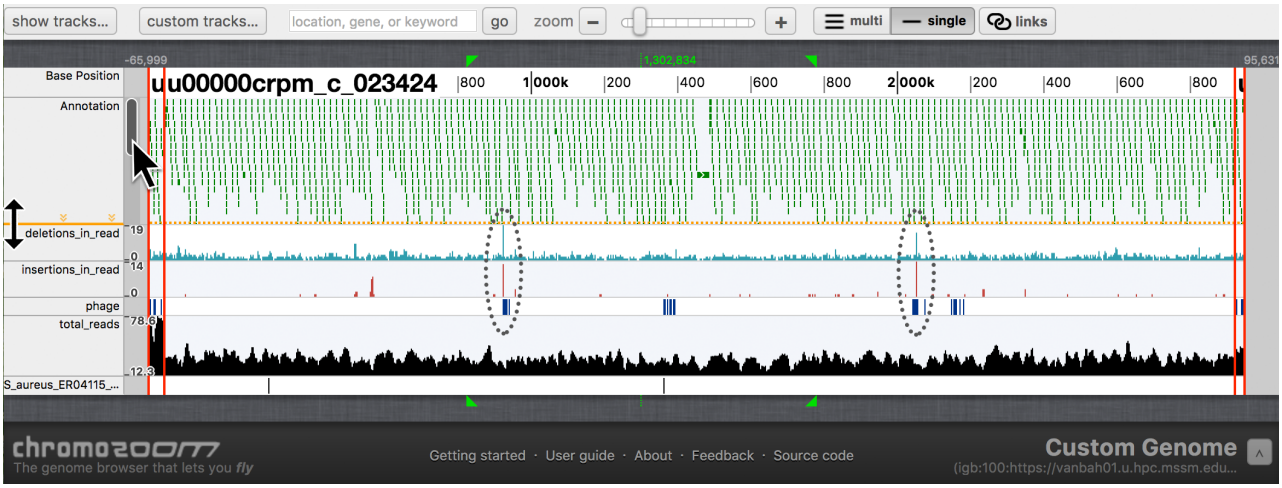


Figure 3.1: **Browsing a completed *S. aureus* assembly with two plasmids** (leftmost and rightmost small contigs). The top track, containing annotations of putative genes, has too much data to display in the available space (as is common for genomic data) which is indicated by the dotted orange line. This can be remedied by resizing the track (double arrow cursor) or scrolling (arrow cursor). Two peaks in the insertions and deletions tracks corresponding to phage regions are highlighted (gray dotted ellipses).

navigational controls, while the genome selection menu is in the lower right. This menu displays "Custom Genome" because the genome layout was not provided by the server, but was loaded directly from another source—in this case, an IGB Quickload Directory. Supported formats for custom genome assemblies will be covered later in this article.

Data is placed front and center in the main area of the interface. The user can click, drag, and throw this area horizontally to move. A green "tank reticle" is visible in the center of the screen, which helps provide feedback during zooming operations with the mousewheel (or two-finger scroll) and highlights the location that the user is zooming on. The display is at nearly the lowest zoom setting, so all three contigs are visible, and are separated by the bright red lines. The names of contigs are shown as the large labels in the "Base Position" track. Ticks in this track follow a minimalist format that avoids duplication of nonsignificant digits.[34]

Six additional tracks are being displayed in this view, which are labeled along the left side of the tracks. The top track, in green, contains gene annotations by `prokka`;[35] this was supplied as a BED track. There are many such annotations in all contigs, and the orange dotted indicator below this track warns the user that some of the data is being clipped by the vertical space available. A scrollbar is available at the leftmost edge of the track to view the bottom edge of the track.

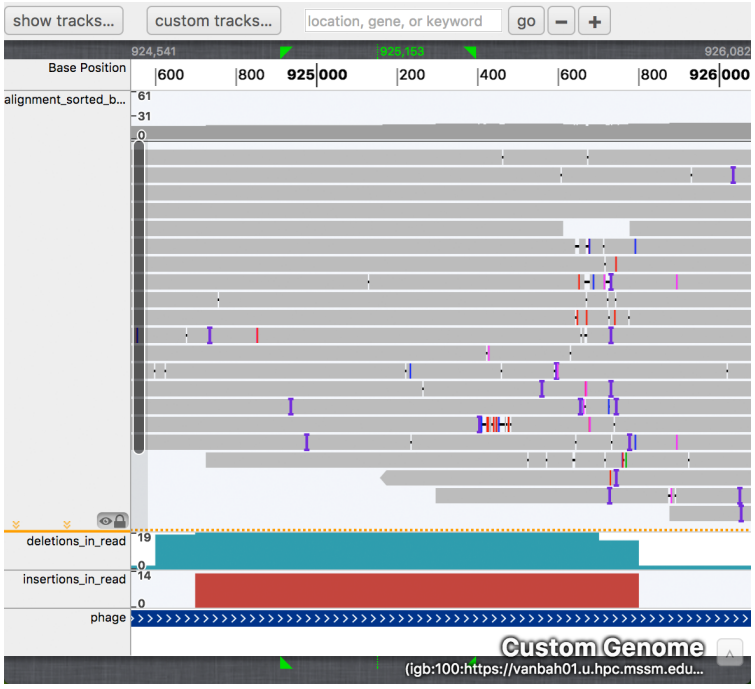
The next two tracks in turquoise and red are bigWig tracks that count the number of deletions or insertions in error-corrected reads aligned back to the final assembly, binned by base position. Two sharp peaks in the main chromo-

[34] Tufte (2001), *The Visual Display of Quantitative Information.* Krzywinski (2013), "Axes, ticks and grids".

[35] Seemann (2014), "Prokka: Rapid prokaryotic genome annotation".

some that reach similar maximum values are notable (see dotted ellipses near the 900k and 2,050k base positions), which happen to correspond with features in the next track that denote putative phage regions.[36] These regions also appear to correlate with troughs in the coverage of aligned error-corrected reads, which is plotted on the subsequent "total_reads" track. The last track shows single nucleotide variants (SNVs) created from a comparison with `nucmer`[37] against a different but closely related hospital strain. Since there are only 2 SNVs separating these strains, they are related enough to consider that transmission (either between the patients or from a common source) was very likely.

A ZOOMED VIEW of the leftmost phage region in the large chromosomal contig is shown in Figure 3.2. (This screenshot highlights the responsiveness of the top and bottom control bars to the resizing of the window: ChromoZoom is usable on displays as small as a smartphone screen.) This confirms that the peak in insertions and deletions is indeed in the middle of an annotated phage region and also a coding region annotated by `prokka`. An even closer look is



depicted in Figure 3.3, which adds an alignment track (in BAM format) of the error-corrected reads created during the Hierarchical Genome Assembly Process.[38] These alignments map the error-corrected reads back to the final contigs

[36] Phages were detected by a custom algorithm written by Mitchell Sullivan and incorporated into pathogendb-pipeline under `scripts/get_repeats_phage_pai.py`

[37] Delcher, Salzberg, and Phillippy (2003), "Using MUMmer to identify similar regions in large sequence sets."

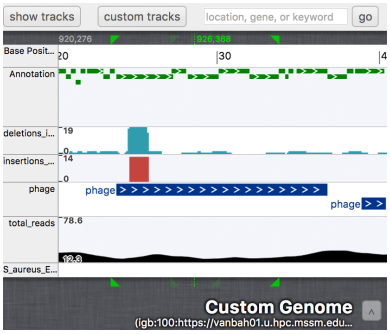


Figure 3.2: A phage region corresponds to high insertion/deletion density in the error-corrected read alignments.

Figure 3.3: **Confirmation of a hypervariable region among phage genes.** An alignment of error-corrected reads to the finished assembly has replaced the gene annotations as the top track, and it uses conventions for displaying BAM files that are similar to IGV. A coverage graph on the top of the track displays the depth of aligned reads, while the alignments themselves are displayed in a stack below, with notation symbols described in the main text. Judging from the totality of the evidence among all reads, the hypervariable region is approximately 200bp long and spans from 925,600 to 925,800.

[38] Chin et al. (2013), "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data."

created during unguided assembly. The track uses display conventions similar to IGV:[39] each aligned preassembled read is depicted as a gray block, with gaps relative to the reference shown as a horizontal black line, and insertions relative to the reference shown as the vertical purple I-bars. Single-base mismatches are depicted as colored vertical stripes in each read. From this view, it appears that an approximately 200bp region is particularly enriched for indels among the majority of the error-corrected reads, and therefore may represent a hypervariable region. Bacteriophages are known to contain genetic elements that generate diversity by site-directed, adenine-specific mutagenesis via reverse transcriptase-mediated exchange between two repeat sequences in order to alter their tropism and possibly also to evade the human immune system.[40] These sequences are relevant to pathogen surveillance because they potentially create inaccuracies in calculating genetic distances between bacterial strains, as they are expected to mutate faster than the remainder of bacterial DNA and therefore should be excluded when constructing phylogenies of hospital strains.

[39] Thorvaldsdóttir, Robinson, and Mesirov (2013).

[40] Liu et al. (2004), "Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements"; Minot et al. (2012), "Hypervariable loci in the human gut virome".

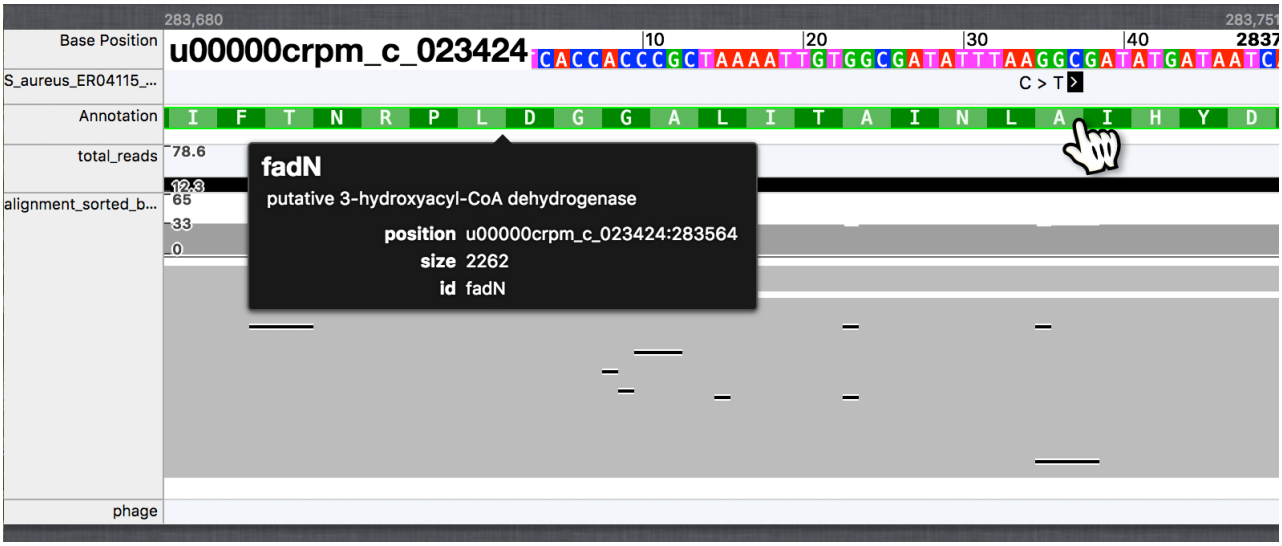


Figure 3.4: A SNV between the two *S. aureus* strains is in *fadN*, which putatively encodes a 3-hydroxyacyl-CoA dehydrogenase. Note that *fadN* is on the negative strand while the SNV annotation (top track) is relative to the positive strand.

When comparing two closely related bacterial strains from different patients or serial isolates from a single patient, as in Chapter 2, it can be revealing to examine the predicted effects of variants in protein-coding regions. These two particular isolates are from different patients and unrelated to a failure in antimicrobial therapy, so a relationship between the SNV and antimicrobial resistance genes would seem unlikely. The right-hand SNV seen in Figure 3.1 is in a

phage region, and is therefore more likely a spurious variant call and unrelated to selective pressure on this strain. The left-hand SNV is not part of a phage region, however, and is shown in greater detail in Figure 3.4. The surrounding region appears to have sufficient coverage by the error-corrected reads and shows generally consistent alignment among those reads.[41] Therefore, this variant call likely reflects a true SNV, although for completeness, the locus for the corresponding SNV in the reverse direction should also be examined. The SNV is in a *fadN* gene, whose homolog was previously identified in *Bacillus subtilis* to be part of a 3-hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase complex, which is involved in fatty acid degradation.[42] This variant, which is aligned to the positive strand while the *fadN* annotation is on the reverse strand, would cause a nonsynonymous Ala-697→Thr mutation. Since both isolates were cultured from separate, active infections, this mutation is most likely not deleterious to fatty acid degradation in *S. aureus*. (A more detailed functional analysis is beyond the scope of this article.)

[41] After error correction of PacBio RSII reads, the most typical remaining errors are short indels, as seen above.

[42] Matsuoka, Hirooka, and Fujita (2007), "Organization and function of the YsiA regulon of Bacillus subtilis involved in fatty acid degradation".

This example shows that ChromoZoom v2 is able to quickly assess the quality and content of new bacterial genome assemblies and variant calls between related strains, which is crucial if infectious disease clinicians expect to use these data as evidence for or against transmission and especially if the timeliness of potential interventions to prevent an outbreak is at stake. Because all data for this visualization was loaded from data accessible via the web, the URL for this view in ChromoZoom can be sent to colleagues to easily share the visualization and its data, who cant then explore it further. This allows labs and clinical groups to collaborate on continuously updated NGS data from local pathogens or any other newly sequenced organisms.

Browsing NGS data for a human genome

ChromoZoom v2 is equally adept at browsing NGS data aligned to standard human and vertebrate references, which is a common task for both research and clinical genetics laboratories.[43] Figure 3.5 depicts a ChromoZoom v2 visualization of NGS reads for the author's genomic DNA generated as part of the Practical Analysis of a Personal Genome class offered at Mount Sinai.[44] Sequencing was performed to approximately 30-fold mean coverage on an Illumina HiSeq using a 100bp paired-end protocol. Any ChromoZoom user can

[43] Human-aligned data is relevant to infectious diseases when finding host biomarkers for infection outcomes; see Chapter 6.

[44] Linderman et al. (2015), "Preparing the next generation of genomicists: a laboratory-style course in medical genomics".

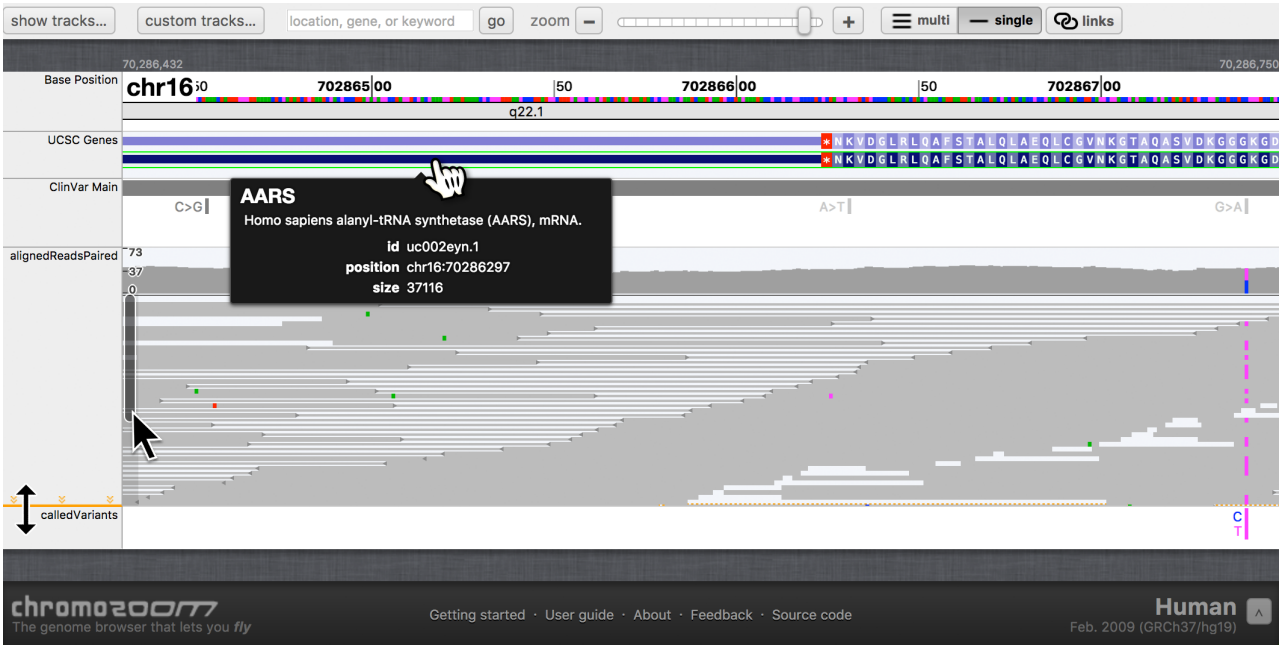


Figure 3.5: **NGS reads for the author's genome aligned to GRCh37/hg19**, displayed in paired mode, and corresponding variant calls. The "UCSC Genes" and "ClinVar Main" tracks are mirrored from UCSC, while the other two were loaded from URLs pasted into the interface by the user; "alignedReadsPaired" is in BAM format while "calledVariants" is a tabix-compressed VCF file. A heterozygous C/T variant in the *AARS* coding sequence is visible to the righthand side, which is annotated in the ClinVar track as a benign variant (light gray G>A; annotated to reverse strand). More alignments are stacked vertically than can be displayed, as indicated by the orange clipping indicator, which could be remedied by resizing the tracks or using the adjacent scrollbar (mouse cursors).

add BAM files (with alignments) and tabix-compressed VCF tracks (containing variants)—both standard products of current NGS pipelines—to the visualization by clicking the "custom tracks..." button in the top toolbar and pasting URLs for the data. Paired-end BAM files can be displayed in both single-end and mate-paired mode.

As described in Implementation, for all of UCSC's active reference genomes, I utilized a track scraper to pull links to all supported track types in UCSC's MySQL database (with conversion to the bigBed format as necessary), which gives ChromoZoom users quick access to a total of 7,969 UCSC-curated tracks for the GRCh37/hg19 reference. Clicking items in these tracks (such as the *AARS* gene in Figure 3.5) takes the user to corresponding item details page in UCSC.

With so many available tracks, searching and selecting interesting annotation sources can be challenging task for the UI to accomodate without overwhelming the user. ChromoZoom v2 uses the streamlined approach of a searchable, expandable list of tracks organized in the same hierarchy that UCSC uses (Figure 3.6). In this UI, I encourage discoverability by limiting the initial list to 11 expandable categories, with no more than 20 tracks and/or track groups inside each category. For the track groups, indicated by the left-hand arrows, the user can "drill down" into subtracks by expanding the hierarchy similar to the

folder-tree views in most file managers. If the user is interested in a particular keyword, like "kidney," they can start typing it, and this pulls all relevant subtracks to the forefront of the hierarchy and also filters the list to only the tracks that at least partially match the query (Figure 3.6). Once the user sees tracks that are potentially interesting, adding the data to the browser view is as simple as clicking a checkbox.

Searching for features within tracks is equally necessary for fast navigation and is depicted in Figure 3.7. For all feature data scraped from UCSC's database, I make use of B-tree indices that can be added to the end of the bigBed formatted tracks that are saved to our server. These indices were added to the bigBed format specification after its first version was published,[45] but can now be easily created using the `-extraIndex` option of the `bedToBigBed` program from Jim Kent's utilities. B-trees allow fast prefix-based searching of prespecified fields in large tracks (even with millions of elements), which I automatically perform on UCSC tracks for fields like feature names and IDs. Searching is as easy as typing in the search box at the top of the browser, which immediately provides suggested results as the user types (similar to autocomplete on smartphones or Google Suggest).[46] Using the search box, users can quickly jump to locations by chromosome coordinates, gene name, or other feature ID. For UCSC genomes, the default gene track is always included in searches, but any other tracks that are currently active will also be automatically searched too (Figure 3.7).

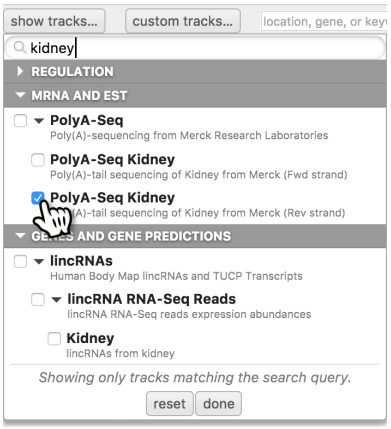


Figure 3.6: A searchable track interface allows quick discovery of relevant tracks from the thousands of tables available from UCSC, each of which can be added with a single click.

[45] Kent et al. (2010).

[46] Google Suggest was rolled out to the homepage of Google in 2008. See https://goo.gl/2053MP

Figure 3.7: On all UCSC reference genomes, the primary gene track will be prefix-searched by name or ID. Also, any visible annotation tracks will be prefix-searched. In this example, the "Fosmid End Pairs" track (which catalogs all valid pairs of fosmid end sequences) was active, and therefore its items are also included in the search results. Results can be navigated by mouse and keyboard. Selecting a result will jump to its position in the browser.

On genomes at the scale of the human reference assemblies, particularly when the data for each track may need to be fetched from different servers, it is crucial for genome browsers to detect and gracefully handle cases when the user is viewing a region that has more data than can be fetched in a reasonable amount of time. Otherwise, as often occurs in other AJAX-using genome browsers, the initiation of a "too-large" data request can stall drawing of data interminably, cause the UI to become unresponsive, or in the worst case, crash the user's web browser. Mature browsers like IGV and UCSC are generally smart enough to handle these cases, e.g., in IGV any region in a BAM file that contains too many reads is automatically downsampled. For ChromoZoom v2, in the case of bigBed, BAM, and tabix compressed formats, I estimate the average feature density when the track is first loaded into the browser, and use that to calculate the optimal width of a range request. This width is then used for

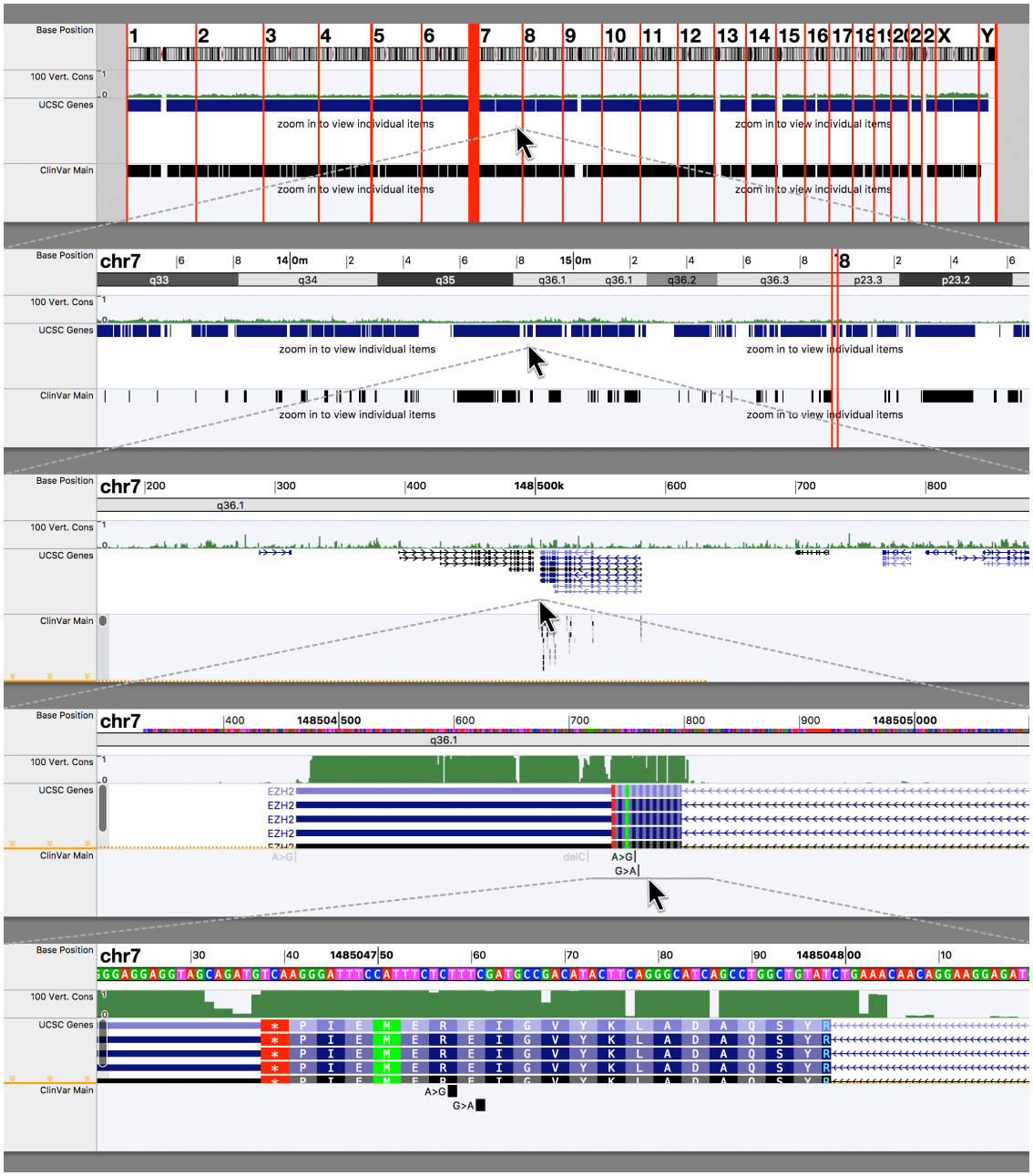


Figure 3.8: **ChromoZoom dynamically redraws ticks, ideograms, and tracks to suit the user's current viewport and zoom level.** The above shows a series of progressive zoom operations on three UCSC-curated tracks for the GRCh37/hg19 reference genome, starting from a view of all chromosomes and contigs and moving in towards a view of two specific variants within the *EZH2* gene. The top track displays base position and ideogram bands, which are automatically overdrawn with nucleotide data as the user zooms. A green bigWig track displaying evolutionary conservation across 100 vertebrates is immediately underneath, followed by UCSC's gene track (dark blue) and the same uncolored ClinVar track seen in Figure 3.5. Mousing over the variants in the bottom view would reveal that they are both pathogenic and associated with Weaver syndrome.

all requests, with returned data cached on the client-side in an intermediate layer (internally, a `RemoteTrack` object) to avoid unnecessary refetching of any region. If the user is attempting to display a region size for which fetching item-level data is impractical, a summary of the data is requested instead, such as the positions covered by at least one item. Figure 3.8 shows this in action for the two lowest zoom levels on the "UCSC Genes" and "ClinVar Main" tracks, which have enough vertical space to start stacking features, but instead inform the user that they should "zoom in to view individual items." As the user passes the estimated threshold for practical viewing of individual data, ChromoZoom automatically begins fetching and drawing them in progessively increasing detail (bottom three rows of Figure 3.8).

This highlights another ability that most current genome browsers lack, which is the seamless adjustment of the draw detail level to best represent the amount of data within the available space. Although in certain cases UCSC and IGV can adjust their drawing styles automatically, in most cases, they expect users to choose a style for each track from the rather obtuse choices of "dense," "squish," and "pack." By and large, this tactic of laying the choice on the user has carried over into next-gen genome browsers.[47] By contrast, Google Maps and other online mapping websites have always applied cartographic principles in condensing and simplifying features to suit the map's scale, so that, e.g., at the world level only country borders and terrain are visible, while at the street level, individual buildings, transit stops, and business labels are visible—and this all occurs automatically while zooming. ChromoZoom adopts this approach, and whenever the user zooms it adjusts its drawing styles to best fit the data to the viewport, including the automatic addition of codons and individual nucleotides at the closest scales (bottom rows, Figure 3.8). If this algorithm isn't providing as much detail as the user needs, perhaps because the user wants to examine or click individual items even when they can't all fit on the screen, the user is always able activate the most detailed display available by using a single button in the sidebar (Figure 3.9).

Most importantly, ChromoZoom takes great care to maintain immersion in the "landscape" of genome features as the user zooms through all levels, using animation of the UI during zoom and redraw operations to keep continuity between all the rendered items even if details are added or removed. These animations can't be depicted in Figure 3.8, but they are central to the design of

[47] igv.js, Buels et al. (2016), Down, Piipari, and Hubbard (2011), and Vanderkam et al. (2016)

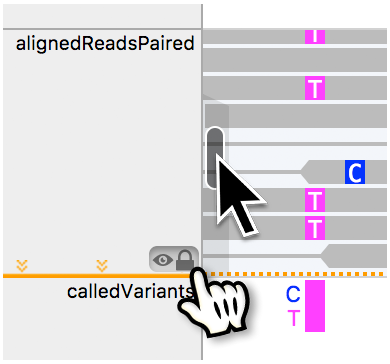


Figure 3.9: If users need to see individual items even if there are too many to fit vertically, there is an "eye-lock" button in the bottom right corner of each track's sidebar (hand cursor). This forces the display of individual items, if possible, and then the user can then scroll vertically as needed with the scrollbar (arrow cursor).

ChromoZoom and therefore are generally achieved with better fidelity than in other web-based genome browsers, which typically have to display spinners or change the layout of onscreen items during jarring pauses.

As I have shown here, ChromoZoom v2 offers powerful new features for the interactive display of NGS data aligned to vertebrate references, and therefore is useful for placing BAM, VCF, and other user-created annotation data in the context of the thousands of standard annotation sources in UCSC's database. As with the *S. aureus* example, views to custom data can be easily shared by URL.

Supported genome formats

Using the genome picker in the bottom right of the screen, users are able to change the currently displayed genome assembly to others available from UCSC, or load their own in several formats supported by ChromoZoom. These formats currently are: FASTA files, which only contain contigs and sequence data in plain text; GenBank files, which contain contigs, sequence data and annotations in plain text; `chrom.sizes` files, which are lists of contigs and their lengths that do not contain any sequence data; and IGB Quickload directories, which can contain a contig layout, sequence data, and both binary and plain-text annotation data. Of these formats, the first three (FASTA, GenBank, and `chrom.sizes`) can be read directly from the user's hard disk, pasted from the clipboard, or fetched via URL.

IGB Quickload Directories cannot be read directly and must be uploaded somewhere on the web for usage with ChromoZoom, so they require additonal setup compared to the other formats. However, they are also the most feature-rich, as the user is able to customize the display of any number of annotation tracks in any format. The *S. aureus* genome displayed in Figures 3.1-3.4 was provided to ChromoZoom as an IGB Quickload Directory, and it contained a mixture of BAM, BED, and bigWig tracks, along with sequence data in the form of a `.2bit` file.[48] As the name suggests, the Quickload format was developed as a way of easily loading annotated genomes into IGB. ChromoZoom retains complete compatibility with IGB's specifications so either program can open the same files, and full instructions for creating a directory are available in IGB's documentation.[49] Briefly, a minimal directory should include a `.2bit`

[48] TwoBit sequence is a standard format created by UCSC and is described at `https://genome.ucsc.edu/goldenpath/help/twoBit.html`. It is a binary format that is more compact than FASTA, and can be created from FASTA files using the `faToTwoBit` executable in Jim Kent's utilities.

[49] A tutorial is available at `https://wiki.transvar.org/display/igbman/Sharing+data+using+QuickLoad+sites`

file with the sequence data, a plaintext `genome.txt` file (that can be generated from the `.2bit` file) listing all contigs and their sizes in the preferred order, and an `annots.xml` file that specifies the annotation tracks that should be loaded alongside the genome.

Supported track formats

A DIVERSE ARRAY of track formats can be loaded into ChromoZoom v2, and due to the new design of loading all data from these files rather than a backend database of cached image tiles, every format is handled equally well whether supplied by the server or by the client. Plaintext formats are the easiest for users to edit by hand or generate from spreadsheet programs, and remain appropriate for data that can be uploaded or downloaded in their entirety in a reasonable amount of time (becoming cumbersome above ~10MB).[50] For interval-based features, such as gene predictions or alignments, ChromoZoom supports the BED format defined by UCSC, including the use of arbitrary fields at the right end of each line (sometimes called BED plus or BED extra fields),[51] and also parses the standard `FEATURES` section of any GenBank files loaded as a custom genome. For continuous quantitative data, both bedGraph and WIG are supported, again using UCSC's specifications.

Binary formats are generally preferable because they allow for compression, indexing, and incremental loading from remote locations via the use of HTTP Range requests,[52] although they do require slightly more setup and must be uploaded somewhere on the web to work with ChromoZoom. ChromoZoom supports bigBed files for interval features, BAM files for alignments, bigWig files for continuous quantitative data, and tabix-compressed VCF files for variant calls. URLs pointing to any of these filetypes can be pasted into ChromoZoom and they will display alongside the currently loaded reference.[53] These are also the same file types now used to display all data for the UCSC reference genomes. bigBed files may include extra fields in additional columns similarly to BED files, which ChromoZoom uses to display expanded tooltips containing customized per-item information that UCSC includes for most of their curated tracks.

Despite our best intentions (and those of users providing custom track data), sometimes things go wrong while adding custom tracks to ChromoZoom. In this case, I try to generate a user-friendly message describing what the error was,

[50] Because plaintext files don't include any indexes, the whole file must be obtained before any data can be displayed.

[51] This requires the use of a track line that lists the extra fields' names as a comma-separated list in a ChromoZoom-specific option called `bedPlusFields`, e.g.,
`track type="bed" bedPlusFields="x,y,z"`

[52] Li et al. (2009); Li (2011); Kent et al. (2010).

[53] The direct upload of these files isn't permitted, because they can easily exceed 100GB apiece and therefore would be impractical to store, even temporarily. I do intend to support opening files directly from users' Dropbox accounts, however—coming soon!

and if possible, where in the file the error was encountered. A sample message is depicted in Figure 3.10.

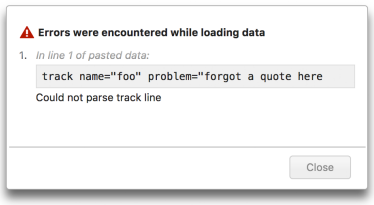


Figure 3.10: If an error occurs while reading a custom track file, ChromoZoom does its best to point out to the user what went wrong, and where in the custom track file things went wrong.

Conclusions

ChromoZoom v2 is the first web-based genome browser that supports client-side loading of custom genome assemblies with rich annotation data. I use this new version of ChromoZoom to assess bacterial assemblies and comparative genomics analyses created as part of pathogen surveillance activities at The Mount Sinai Hospital. However, the new release of ChromoZoom is equally capable of loading standard references for vertebrate genomes and displaying the results of typical NGS experiments, such as alignments of paired-end reads from an Illumina instrument, alongside thousands of curated tracks from UCSC. I contend that ChromoZoom v2 is the genome browser that best achieves the three core principles of the user experience championed by Google Maps, in that it is (1) universally accessible, (2) immersively interactive, and (3) easily sharable. Analysis of NGS data continues to be an increasingly collaborative endeavor, with participants from diverse backgrounds (spanning bioinformatics, basic science, and clinical disciplines) who are also often spread across multiple institutions. I envision ChromoZoom as the premiere genome browsing platform for team-oriented, interactive visualization of NGS data.

Notes

Contributions

Theodore R. Pak (trp), Miha Skalic (ms), Adrian Pasculescu (ap), and Frederick "Fritz" P. Roth (fpr) contributed to this chapter. trp wrote the first and second versions of ChromoZoom. ms and ap contributed to the Python pipeline for mirroring data from UCSC Genome Browser. fpr provided advice and support during the initial development of ChromoZoom and throughout the development of the second version. trp wrote the first draft of this chapter. trp is the first author on the corresponding manuscript.

Funding

TRP was supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai and NIH/NIAID (U19-AI118610 and F30-AI122673). Research was also supported by the Office of Research Infrastructure of the NIH under award number S10-OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

I thank Andrew Kasarskis for supporting this project and Miguel Ranjo for providing bug reports. This work was supported in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

4

The PathogenDB software suite for genomic clinical microbiology & epidemiology

Next-generation sequencing (NGS) technologies have reduced the cost of acquiring genomic data from active infections in hospitals, with the potential to rapidly characterize patient-to-patient transmission with extreme precision. However, there is no integrated software for converting NGS data into species identifications, phylogenies, and drug susceptibilities, with particularly few options including de novo *assembly. A clinical application would ideally provide a unified pipeline that runs semi-automated analyses to inform infection prevention and control interventions. We developed a modular open-source software suite called PathogenDB that implements major functionalities needed for genomic clinical microbiology and pathogen surveillance. A central laboratory information management system runs on a standard open-source Linux/Apache/MySQL/PHP stack. A modular genomics workflow, PathogenDB-pipeline, automates de novo assembly, circularization, gene annotation, quality control, and epigenetic motif prediction. A comparative genomics module, PathogenDB-comparison, performs semi-automated phylogenetic analysis. Finally, a visualization suite, PathogenDB-viz, integrates phylogenies and epidemiological data into a "live view" of putative transmissions mapped to hospital locations. Thus far, PathogenDB-pipeline has been used to assemble and annotate 593 genomes from 7 species, and runs in <12 hours end-to-end. At The Mount Sinai Hospital, PathogenDB-comparison has genomically characterized one MRSA outbreak, two transmissions via solid organ transplant, and pseudo-outbreaks of* S. maltophilia *and* B. cepacia. *All three software packages are freely available on GitHub.*

JOHN HAMMOND: *Dennis, our lives are in your hands and you have butterfingers?*

DENNIS NEDRY: [laughs] *I am totally unappreciated in my time. You can run this whole park from this room with minimal staff for up to three days. You think that kind of automation is easy? Or cheap? You know anybody who can network eight connection machines and debug two million lines of code for what I bid for this job?*

—Jurassic Park

Let me put it this way, Mr. Amor. The 9000 series is the most reliable computer ever made. No 9000 computer has ever made a mistake or distorted information. We are all, by any practical definition of the words, foolproof and incapable of error.

—HAL 9000, *2001: A Space Odyssey*

THERE IS increasing consensus that next-generation sequencing (NGS) technologies will eventually become mainstream equipment in clinical microbiology laboratories,[1] considering that its nominal reagent cost is already within

[1] Didelot et al. (2012), "Transforming clinical microbiology with bacterial genome sequencing"; Harris et al. (2013), "Whole-genome sequencing for analysis of an outbreak of methicillin-resistant Staphylococcus aureus: a descriptive study."; Joensen et al. (2014), "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli."

range of routine tests ($25 per isolate for certain short-read technologies) and that it is likely to improve turnaround time and sample throughput for epidemiological investigations and hard-to-culture organisms.[2] One significant barrier to this, as noted in Chapter 1, is that informatics and software infrastructure for the new diagnostic workflows afforded by these technologies are not yet widely available. While robotic culturing systems like Vitek and BD Phoenix include mature software packages for automating the executing and interpretation of routine tests, even integrating directly with standard laboratory information management systems (LIMS) so that results can be associated with patient metadata and sent directly back to ordering physicians via the electronic medical record (EMR), no such framework exists for genomic clinical microbiology.

We have already noted in Tables 1.1 and 1.2 and in Chapter 1 that many open source software packages exist for *individual components* of such a pipeline but no end-to-end solution has yet been assembled by the research community. The most mature components that are currently available are typically steps closest to the sequencer, since the relatively small number of sequencing platforms and ubiquitous demand for certain invariant processing steps for their direct outputs (e.g., debarcoding and demultiplexing, filtering reads, quality control, alignment to a reference) have spurred researchers to create consensus solutions for them. Less mature are the steps involved in *de novo* assembly and beyond this frontier, since the capability to finish assemblies without human intervention has only recently emerged[3] and the underlying algorithms and heuristics are still an area of active research.[4] Finishing, annotating, and comparing brand new bacterial assemblies, therefore, has been sufficiently niche that a standard distributable "toolkit" was not needed. However, as long-read sequencing continues to drop in price and complexity[5] and *de novo* assembly for these platforms becomes more accessible, demand for pathways bringing these data into medical microbiology is bound to rise.[6]

Fortunately, the construction of new pipelines built around existing smaller tools is a typical task in bioinformatics—so common, in fact, that meta-tools are available for this very purpose.[7] The fact that a standard pipeline has been slow to emerge should not be considered a bad omen; in fact, this same situation existed for human genomes when short-read NGS technologies first emerged. As large-scale community resource projects like 1000 Genomes were launched to take advantage of NGS, efforts to develop standardized pipelines to execute

[2] Didelot et al. (2012); Köser et al. (2012), "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology."

[3] Bashir et al. (2012), "A hybrid approach for the automated finishing of bacterial genomes".

[4] Sohn and Nam (2016), "The present and future of de novo whole-genome assembly".

[5] The MinION, by Oxford Nanopore Technologies, is available in $1,000 starter kits, plugs into a USB port, and produces reads up to 10kbp; see Check Hayden (2014)

[6] Reuter et al. (2016), "Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology".

[7] Köster and Rahmann (2012), "Snakemake-a scalable bioinformatics workflow engine"; Goecks, Nekrutenko, and Taylor (2010), "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences."; Jamil (2013), "Designing integrated computational biology pipelines visually."

those projects naturally followed and bore fruit. The landmark publication for 1000 Genomes was filed in 2010,[8] and within the few short years that followed, the tools that were developed to enable the project's goals were released,[9] refined into best practices for the community,[10] and reached sufficient maturity that they could be validated and adopted for use in clinical whole genome and whole exome sequencing pipelines.[11] These tools, particularly Broad's Genome Analysis Toolkit (GATK), took what had previously been a hodgepodge of nascent data formats and algorithms and unified them into a consistently designed library with a uniform API. Furthermore, they were optimized for re-use in diverse and parallelized computing environments, and most importantly underwent battle-testing in enough real-world use cases to become de-facto standards. Given the rapid pace at which this occurred, we remain optimistic that similar widescale efforts in genomic microbiology can drive demand and community support for a similar toolkit for pathogens, and that such software's adoption can reach clinical laboratories within years of its initial release.[12]

[8] The 1000 Genomes Project Consortium (2010), "A map of human genome variation from population-scale sequencing."

[9] McKenna et al. (2010), "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data".

[10] Van der Auwera et al. (2013), "From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline".

[11] Linderman et al. (2014), "Analytical validation of whole exome and whole genome sequencing for clinical applications".

[12] Pak and Kasarskis (2015), "How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management".

Of course, there are many challenges that can be expected, although they generally fall into the category of engineering problems and can therefore leverage many of the lessons and algorithms already unearthed by human genomicists. The premise of a genomic microbiology pipeline gaining clinical use requires it to be fast, if not faster than substitutable lab tests like culturing.[13] Since this implies a turnaround time of one or two days, the software must be efficient. Ideally, it would not require the supercomputing power typically associated with human genomics, and could instead run on a single server or a desktop machine that any microbiology lab could afford.[14] Thankfully, bacterial genomic data are generally much smaller than their human counterparts, and many of the components in Table 1.2 were already developed for use on everyday desktop hardware. The software must be generally reliable, but seeing as no bioinformatics pipeline succeeds in real-world usage 100% of the time (imagine contaminated DNA entering an analysis, or the computer running out of disk space), if it must fail, it should fail obviously and provide some clues as to where and why. If possible, it should save work up to the failed step so when analysis needs to be re-run, time does is not wasted re-running steps that previously succeeded. Perhaps most importantly, all steps need to be reproducible, because it would be extremely difficult (if not untenable on its face) to achieve diagnostic validity with a nondeterministic and therefore unauditable procedure. While this

[13] Köser et al. (2012).

[14] Alternatively, labs could "rent" supercomputing power from cloud platforms like Amazon's Elastic Compute Cloud, but this introduces special complexities of its own.

seems simple in principle for a computer program, in practice, with constantly changing sources of "truth" (such as changes in local and remote databases, updates to ancillary libraries, and evolving storage formats) it can be a devilishly complicated affair.

There are some recent examples of publicly released, self-contained pipelines that use NGS to solve specific problems in clinical microbiology. One is SURPI, which processes millions of reads from a metagenomic sample to rapidly search for evidence of pathogen DNA.[15] Most notably, this pipeline was used to deliver a timely diagnosis for a case of neuroleptospirosis that eluded traditional diagnostic assays.[16] It can be deployed on both standalone servers and cloud computing environments.[17] Another is Mykrobe, which similarly processes NGS reads to generate clinican-friendly reports of antimicrobial resistance predictions for *Staphylococcus aureus* and *Mycobacterium tuberculosis*, matching the results of gold-standard methods in 99% of drug-strain combinations for the former and >80% for the latter.[18] Both of these are admirable for being self-contained packages that anybody can download and run (Mykrobe can even run on consumer-grade macOS and Windows laptops) and requiring only one input: raw NGS reads in the common FASTQ format. While third party comparisons have yet to be published, both claim to provide reliable and actionable data. Neither, however, performs epidemiological analysis, i.e., assessment of multiple samples for the likelihood of transmission.

Our goal was to create a software suite to support the aims of the Pathogen Surveillance Program at The Mount Sinai Hospital, which applies long-read and short-read NGS technologies to routinely collected clinical microbiology specimens and aims to track and prevent the spread of healthcare associated infections (HAIs) throughout our health system. As this is a much broader goal than attempted by the aforementioned software packages, we divided our design into four modular, coordinated components: a LIMS suitable for genomic clinical microbiology (PathogenDB), an all-purpose bacterial assembly and annotation pipeline for the PacBio RS II (PathogenDB-pipeline), a comparative genomics toolkit for the outputs of that pipeline (PathogenDB-comparison), and finally a visualization platform to turn these data into something clinically actionable (PathogenDB-viz). In this chapter, we present the implementation of each of these components, the results enabled by the entire software suite to date, and our plan to disseminate the tools for broader use.

[15] Naccache et al. (2014), "A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples".

[16] Wilson et al. (2014), "Actionable diagnosis of neuroleptospirosis by next-generation sequencing."

[17] Naccache et al. (2014).

[18] Bradley et al. (2015), "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis".

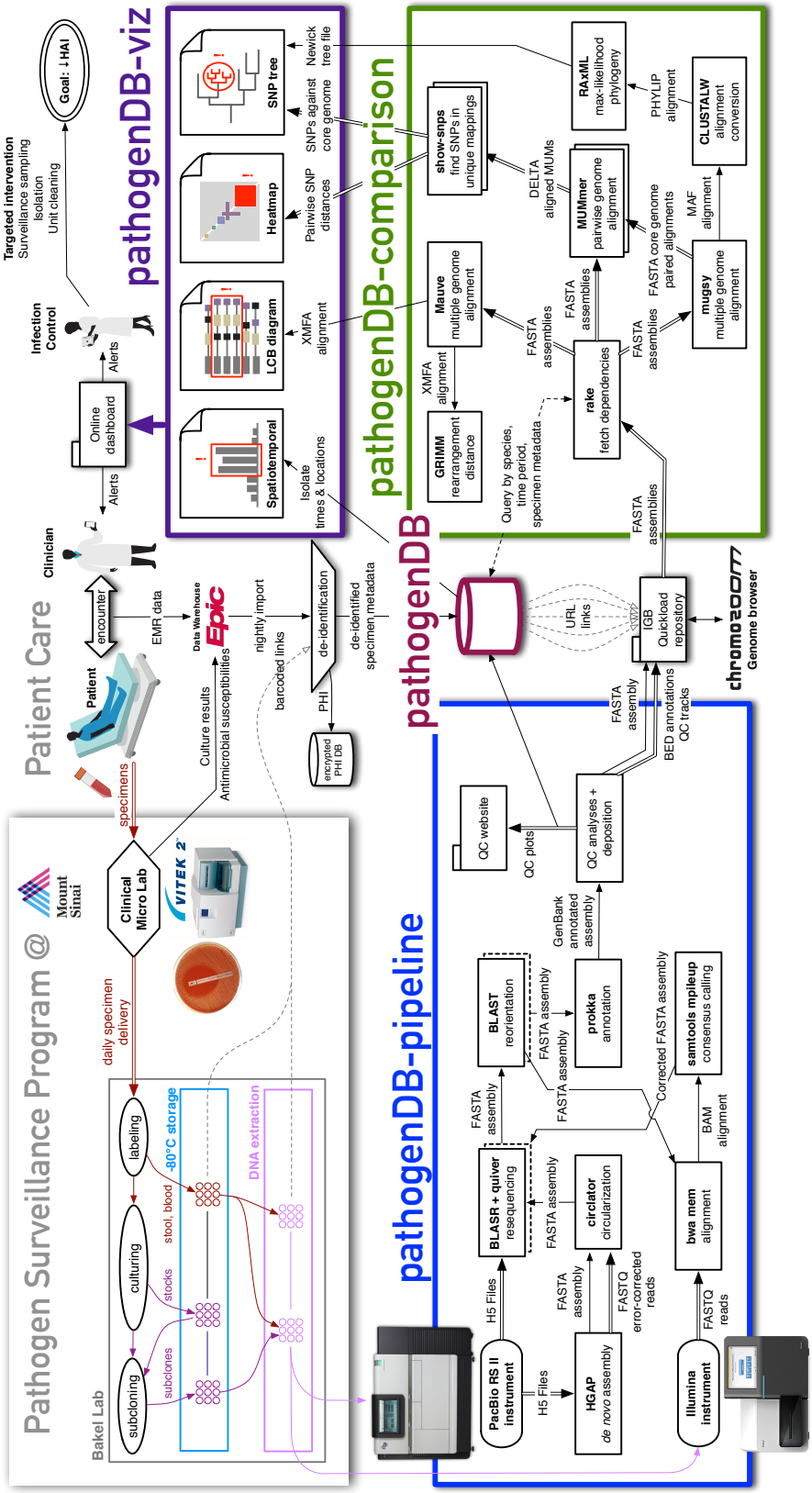


Figure 4.1: **Overview of the PathogenDB suite**. The three modular processing components, PathogenDB-pipeline, PathogenDB-comparison, and PathogenDB-viz, surround our custom LIMS, PathogenDB, which serves as the central "source of truth." The design of these components (and the circular path of information) directly reflects the workflow proposed in Figure 1.2—note that the direction of flow in this diagram is generally counter-clockwise while it is clockwise in Figure 1.2.

Implementation

A TOP-LEVEL VIEW of all steps and their relationships is presented in Figure 4.1. We will review the details of each major component (boxed segments) in sequence. Note that our current design and separation of concerns directly reflects the large circular workflow of a "learning health system" for infectious diseases proposed earlier in Figure 1.2.

Sample collection

The first step of any genomic surveillance project is to collect and organize samples, which is reflected in the workflow at the upper left of Figure 1.2. This is performed by people, not software. In the case of the Pathogen Surveillance Program, the Bakel lab receives daily deliveries of specimens from Mount Sinai's clinical microbiology laboratory. Staff in this lab must carry out a meticulous process of labeling, culturing, stocking, subcloning, and extracting DNA before any sequencing can occur.

However, to track the specimens, stocks, and derivative samples and associate them with metadata created during patient care (top center of Figure 1.2), a database is necessary. In our workflow, we call this database PathogenDB, and it serves as the critical central "source of truth" for all operations and analyses. PathogenDB receives updates via online input forms from the staff in the Bakel lab as they process samples. All items are barcoded and recorded in PathogenDB. The barcoding process, which involves assigning a new ID, both serves to unqiuely identify every item and to remove any pre-existing association with patient identifiers. Simultaneously, PathogenDB receives automatic nightly reports from the EMR (Epic Systems) which contain all the isolates that were expected to be sent for that day along with clinical metadata like the hospital unit of collection, the collection date and time, the bodily source of the specimen (blood, stool, wound, etc.), and an opaque patient ID that does not correspond to any patient IDs reflected in the actual medical record (like the medical record number). Only authorized staff, such as clinicians and infection prevention and control officers, are allowed to see the key linking patient metadata in PathogenDB to outside medical records. This is done to ensure de-identification of Protected Health Information in accordance with HIPAA Safe

Harbor Method principles.[19]

[19] See 45 CFR §164.514(b)(2).

PathogenDB is implemented as a MySQL database, and the relational structure for core tables is depicted in Figure 4.2. The structure of the database

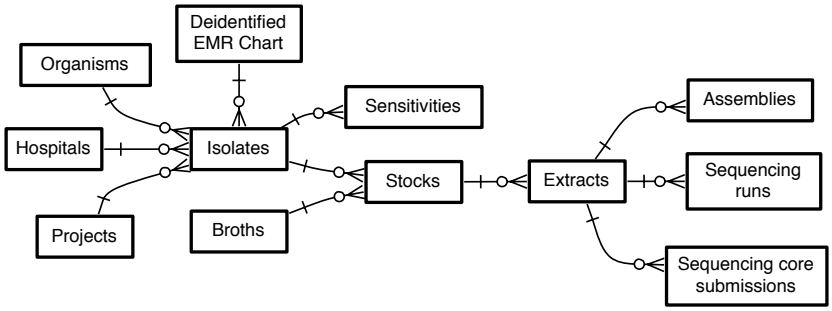


Figure 4.2: **Entity-relationship diagram for the database underlying PathogenDB,** using Information Engineering notation; see Halpin and Morgan (2010). Boxes represent tables of entities; single lines represent relationships, with arrowheads indicating the cardinality of each side of the relationship; crow's foot arrowhead with circle represents "zero or more;" cross-stroke arrowhead represents "exactly one."

mirrors the workflow of banking and preparing isolates, with steps proceeding roughly left to right starting from "Isolates"—i.e., isolates can associated with one or more derivative stocks once culturing and banking are performed, which will then be associated with one or more extracts, which are then submitted to the Genomics Core Facilty at Mount Sinai ("sequencing core submissions"). The returned outputs from Genomics Core are logged as "sequencing runs" and then eventually "assemblies" are created upon completion of the PathogenDB-pipeline (see next section). Maintaining database tables for each step of the process ensures that items are not lost along the way, and that the provenance of every downstream product can be traced backwards, which is crucial if contamination or mishandling are suspected.

PathogenDB provides a frontend that is based on phpMyEdit,[20] which is displayed in Figure 4.3. phpMyEdit allows for quick scaffolding of basic create-read-update-delete (CRUD) webpages for each of the tables in PathogenDB, which can be sorted, searched, edited, and downloaded by authorized members of the team. Authorization can be granted only to particular pages and views so that Bakel lab staff see only tables related to the isolate culturing and extraction workflow, while clinical coordinators instead see tables on patients due for sample collection and forms for sample submission.

[20] `http://www.phpmyedit.org/`

Figure 4.3A shows a sample table view for the Isolates table, which tracks every biosample received by the Pathogen Surveillance Program. Isolates are associated with a Projects; if collected as part of routine operations, this is "CML surveillance," but this field accommodates annotation of samples acquired from

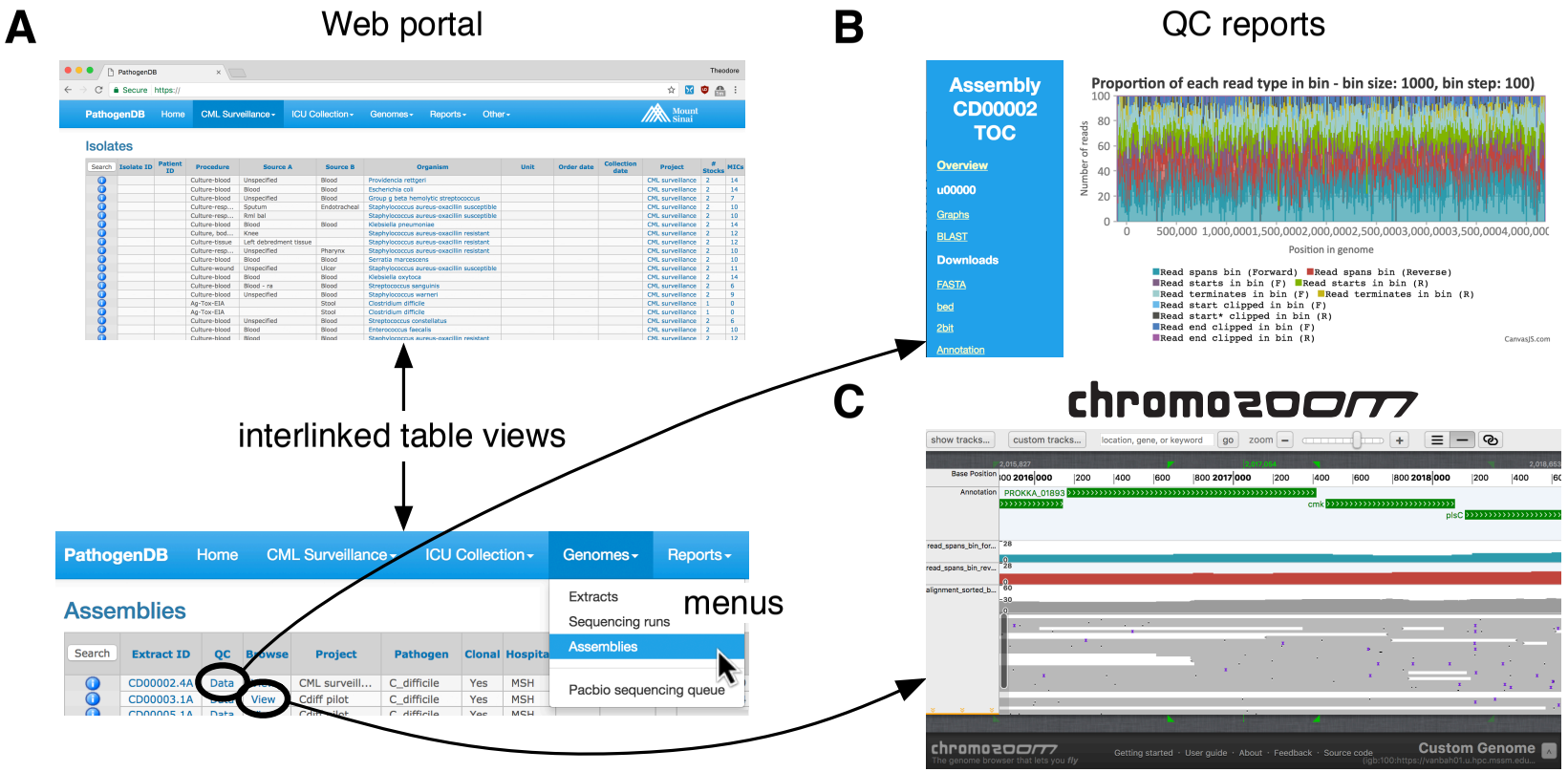


Figure 4.3: **Overview of the web frontend for PathogenDB.** A, the web portal permits quick access to basic views for all tables in the database, which can be searched, sorted, edited, and downloaded. At top, the Isolates table is shown, with certain potentially identifying information removed from the screenshot. At bottom, a zoomed view of the Assemblies table is shown, with the menu for switching between tables also shown. Links to special views (in B and C) are highlighted. B, sample quality control (QC) report for an assembly. Here, a graph of proportions of aligned reads that match various criteria is shown. Extreme fluctuations in these values can indicate assembly problems. C, ChromoZoom displaying a finished assembly, with annotated genes at top, two QC tracks in the center, and alignments of error corrected reads to the finished assembly at bottom. For more on ChromoZoom, see Chapter 3.

ad-hoc and outside investigations. Links at the far right of the table for Stocks and Minimum Inhibitory Concentrations (MICs) indicate that most Isolates are associated with two stocks and up to 14 measurements of antimicrobial susceptibility, which are extracted from Vitek (bioMérieux) results in the reports sent nightly by the EMR. Clicking on these links moves the user to a view of the corresponding entries in the related tables. On the Assemblies page, shown at the bottom of this panel, special views are provided for viewing outputs of the PathogenDB-pipeline (after assembly and annotation are complete). The first of these is a quality control (QC) report (Figure 4.3B), which shows plots of the final contig layout and read statistics, which can be useful for assessing trustworthiness of the assembly and diagnosing reasons for failure to circularize or high fragmentation. The second of these is a link to a ChromoZoom visualization of the genome (Figure 4.3C, also see Chapter 3).

PathogenDB-pipeline: Assembly and annotation from long reads

Once read data are available from the Genomics Core, computational analysis can begin. To date, the Pathogen Surveillance Program has chosen to sequence essentially all isolates on the PacBio RS II (see Methods, Chapter 2). This sequencing platform comes with a manufacturer-maintained analysis toolkit called SMRT-Analysis[21] that we leverage for certain steps, which even includes a web interface (SMRT Portal), but we focus here on a fully automated solution.

[21] https://github.com/PacificBiosciences/SMRT-Analysis

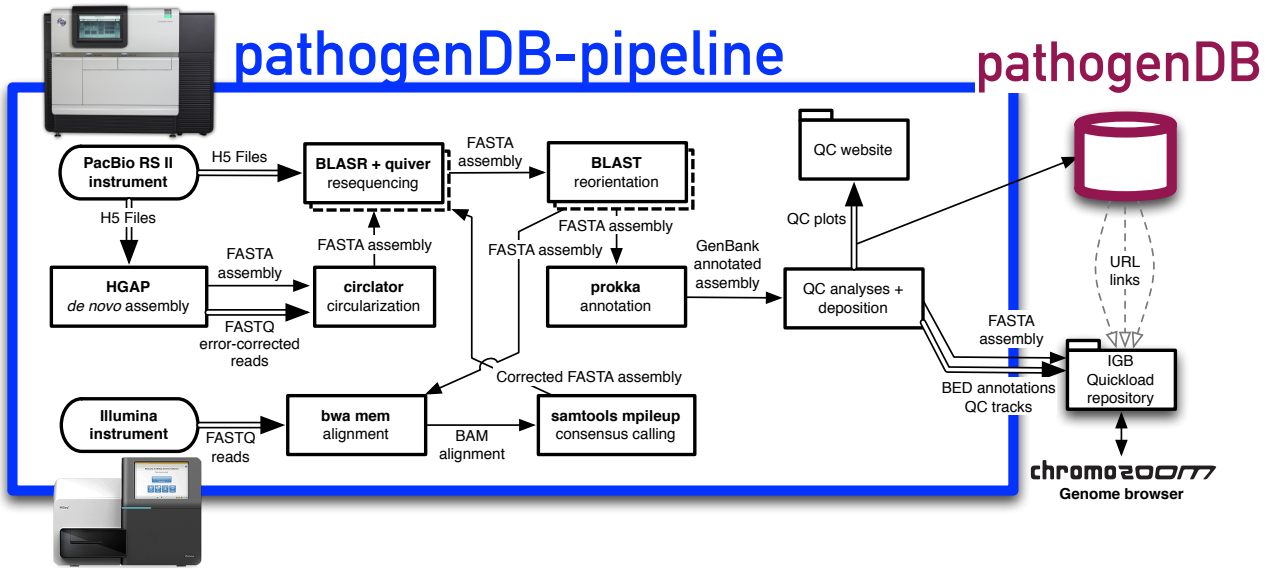


Figure 4.4: **Outline of steps automated by PathogenDB-pipeline.** Processes are depicted as boxes, with processes requiring potentially multiple runs indicated as a "stack." An interim file format is depicted as a single arrow, and groups of files as doubled arrows. The pipeline concludes with deposition of a link to an IGB Quickload Directory for the completed assembly into PathogenDB.

The chosen strategy for converting PacBio RS II reads into a final annotated assembly is outlined in Figure 4.4. Briefly, the H5 outputs of the instrument (which contains movies of the single molecule reads) are assembled *de novo* using the Hierarchical Genome Assembly Process[22] (HGAP), which is included in SMRT-Analysis. This produces a draft assembly, but since HGAP cannot create circular contigs, a FASTA of this assembly and the FASTQs of error-corrected reads produced by HGAP are passed off to `circlator`,[23] a recently released tool that automates *circularization* by using SPAdes[24] to re-assemble the error-corrected reads that overlapped the ends of contigs. This produces a new assembly that is then *polished* over the circularized junctions by re-mapping raw reads using BLASR[25] and recalling the consensus with Quiver,[26] which reduces errors in these regions by re-incorporating read information that could not have aligned properly during the initial HGAP assembly. The polished assembly is then *reoriented* back to the origin of replication (a convention for GenBank bacterial chromosome sequences) using a custom script[27] that performs a BLAST against `circlator`'s suggested origin point for each circular contig, which it decides based on a PROmer[28] search for *dnaA* sequences.

The circularized, polished, and reoriented assembly is finally ready for annotation. Firstly, the contigs are renamed from the overly verbose HGAP and Quiver defaults. We use an in-house convention of starting all contig names with "u" (for "unitig", a term from Celera[29]) followed by a five-digit unitig number originally assigned by HGAP. This is followed by three letters that flag for circularization, reorientation, polishing, and an additional letter reserved for later use, with "x" indicating failure of that step. Another letter surrounded by underscores signals the hypothesized type of contig (chromosome, plasmid, merged, garbage, or other).[30] Finally, the original SMRT-analysis job number set by the Genomics Core is appended. This renaming results in a contig ID like `u00011crpx_p_023011`, indicating it is the 11th unitig, was circularized, reoriented, and polished, is probably a plasmid, and came from sequencing job #023011. These names are short enough for easy viewing in downstream tools like ChromoZoom, while retaining as much information as possible about the provenance and assembly status of the contig. Renamed contigs are finally annotated with `prokka`,[31] which detects putative coding regions and maps them to annotated gene names in UniProt.[32] We then run a series of custom scripts to generate diagnostic files for the QC webpage (Figure 4.3B) and to convert

[22] Chin et al. (2013), "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data."

[23] Hunt et al. (2015), "Circlator: automated circularization of genome assemblies using long sequencing reads".

[24] Bankevich et al. (2012), "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing".

[25] Chaisson and Tesler (2012), "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory."

[26] Chin et al. (2013).

[27] `scripts/post_quiver_orient_correct.py` in the GitHub repo.

[28] Kurtz et al. (2004), "Versatile and open software for comparing large genomes."

[29] See `http://wgs-assembler.sourceforge.net/wiki/index.php/Celera_Assembler_Terminology`

[30] We use the simple heuristic that any successfully circularized contig >1Mbp is a chromosome, and anything smaller is a plasmid. Merged and garbage contigs are the result of ambiguities during assembly. For more detail, see `scripts/post_circlator_contig_rename.py`.

[31] Seemann (2014), "Prokka: Rapid prokaryotic genome annotation".

[32] Wasmuth and Lima (2016), "UniProt: the universal protein knowledgebase".

the assembly and related tracks into an IGB Quickload Directory that can be loaded into ChromoZoom (Figure 4.3C).

PathogenDB-pipeline is implemented as a `Rakefile`, which is written in Ruby and executed by `rake`, an analog of GNU `make`.[33] GNU `make` was originally written as a build system for automating the generation of executables and other products from a program's source files. The advantage of a build system is that it encourages the explicit annotation of dependencies between interim files and tasks into the pipeline, thereby allowing for previous products of a partial build to be automatically re-used if their dependencies (source files) have not changed. From the user's perspective, another benefit is that only the desired final task in the pipeline needs to be specified, and `rake` can automatically figure out what preceding tasks are required to get to that point. For most runs, the sequence of tasks selected by PathogenDB-pipeline is the following:

[33] GNU `make` also inspired SnakeMake, a bioinformatics-focused build system for Python; see Köster and Rahmann (2012).

1. `pull_down_raw_reads`
2. `assemble_raw_reads`
3. `run_circlator`
4. `post_circlator`
5. `resequence_assembly`
6. `post_quiver_orient_correct`
7. `prokka_annotate`
8. `create_QC_webpage`
9. `prokka_to_igb`

which mostly correspond, unsurprisingly, to the boxes in Figure 4.4. Most of these tasks require at least one configuration option, such as the expected species or the destination for output. These are specified as environment variables during invocation, so to get to the final `prokka_to_igb` step above, the user would run something like:

```
$ rake OUT=scratch/out/ER05681 \
       SMRT_JOB_ID=023154 \
       STRAIN_NAME=ER05681 \
       SPECIES="Staphylococcus aureus" \
       prokka_to_igb
```

If problems occur during assembly, the pipeline supports manual editing

of the interim FASTA files, and then a flag (`CURATED=1`) can be set to signify that manual curation occurred and to therefore bypass circularization, reorientation, and contig renaming. There is also an optional branch (bottom entry point for Figure 4.4) that can incorporate Illumina short reads during assembly. Auxiliary sequencing on a short read platform can correct small errors that HGAP on PacBio reads will miss—typically indels in homopolymeric stretches. For this branch, `bwa`[34] is used to perform in-memory alignment against the circularized, reoriented, and polished assembly, and `samtools`[35] and `vcftools`[36] are used to call variants from the pileup and create a new FASTA consensus. These steps can in fact be repeated several times (crossover loop in Figure 4.4) to call a progressively more accurate consensus, although this repetition must currently be invoked manually.[37]

[34] Li and Durbin (2010), "Fast and accurate long-read alignment with Burrows-Wheeler transform".

[35] Li et al. (2009), "The Sequence Alignment/Map format and SAMtools".

[36] Danecek et al. (2011), "The variant call format and VCFtools".

[37] A future version of the pipeline might attempt to re-run the steps until the consensus stabilizes.

The modularity of siloing tasks in a `Rakefile` provided long-term advantages besides those mentioned previously. In our case, when we first created the pipeline in 2013, mature tools for some of the steps did not yet exist, e.g., `circlator` and `prokka` were not yet publicly available. Therefore, we used less efficient solutions, such as our own custom script for circularization and the RAST web service[38] for annotation, as in Chapter 2. Once mature tools became available, it was relatively simple to swap them into the pipeline while preserving the old code under an `old:` namespace, indicating tasks that are deprecated. Because each task and its dependencies are relatively isolated, different members of the team can create slightly different versions of certain interim steps while still sharing all code in one common `Rakefile` pipeline. This was of course further reinforced by keeping all code under version control.[39]

[38] Aziz et al. (2008), "The RAST Server: rapid annotations using subsystems technology."

[39] `https://github.com/powerpak/pathogendb-pipeline`

PathogenDB-comparison: Rapid comparative genomics

After deposition of a complete annotated assembly's IGB Quickload Directory and corresponding record into the Assemblies table of PathogenDB, the next logical step of analysis is to compare all genomes for a given species (perhaps with some filtering by timeframe or location) to determine relatedness and the likelihood of transmissions. We implemented these analyses within the next module of our suite, called PathogenDB-comparison.

Our implementation of workflows for this stage of analysis is outlined in Figure 4.5. As emphasized previously, PathogenDB is considered the single

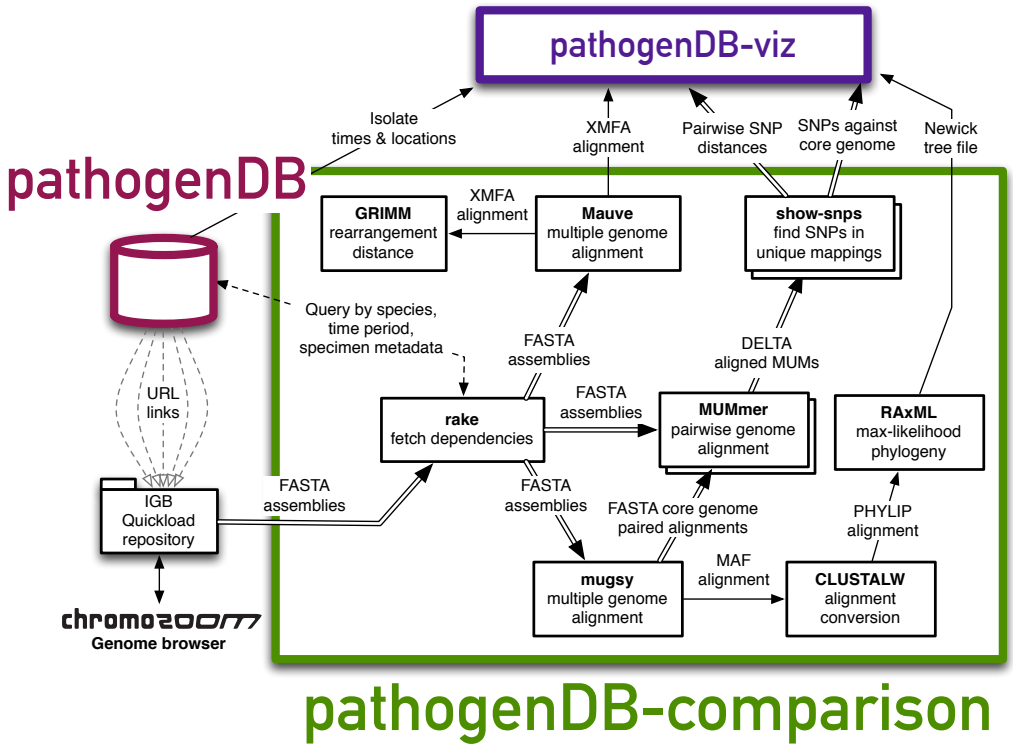


Figure 4.5: **Outline of steps automated by PathogenDB-comparison.** Processes are depicted as boxes, with processes requiring potentially multiple runs indicated as a "stack." An interim file format is depicted as a single arrow, and groups of files as doubled arrows. The pipeline concludes with various outputs being sent to PathogenDB-viz for further visualization.

source of "truth" from which all assemblies and metadata are queried before running an analysis; however, some of the tasks are generic enough to run on an arbitrary set of FASTA files without metadata. Like PathogenDB-pipeline, PathogenDB-comparison is also implemented using `rake`, but its workflow is more branched. The three types of implemented analyses, reflected in Figure 4.5 by the three arrows emerging from the `rake` box and listed here with their corresponding `rake` task names, are:

1. `mauve`: Mauve alignment, which highlights structural variants
2. `snv`: Pairwise MUMmer single nucleotide variant (SNV) distances for heatmap visualization
3. `mugsy`: Core genome alignment for a phylogeny with branch lengths scaled to SNV distances

A Mauve alignment of *S. maltophila* genomes was previously depicted in Figure 2.4 and is most useful for finding large insertions, deletions, translocations, and other recombinatorial events. Mauve performs alignments by using an anchoring heuristic to search for large areas of homology among subsets of the input genomes, which it terms local collinearity blocks (LCBs).[40] Our correspond-

[40] Darling, Mau, and Perna (2010), "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement".

ing task simply wraps execution of `progressiveMauve`[41] and returns the XMFA alignment, since visualization is typically performed with the Mauve Java application. However, we also provide a task that can calculate pairwise rearrangement distances using GRIMM,[42] which searches for the minimal number of inversion operations needed to transform one genome's sequence of LCBs into another genome's. Because inversion distance is algorithmically simple to calculate[43] but probably reflects evolutionary edit distances less accurately than newer models like double-cut and join (DCJ),[44] we have not yet made full use of these distances, but hope to eventually incorporate a wide array of structural variant edit distances into downstream analysis.[45]

MUMmer is a versatile software suite for fast pairwise genome alignment that finds maximally exact matches (formerly maximal unique matches, hence MUM) using a suffix tree algorithm that can run in linear time.[46] Although it is excellent for finding subsequences of one genome that are within a certain edit distance of all locations on another genome, because of the strict edit distance threshold, it is less suited for finding large rearrangements compared to Mauve. However, it is very well suited for quickly calling SNVs between two genomes, as long as the SNVs are not so closely spaced as to elude a maximally exact match for the surrounding region (which should occur only extremely rarely). For this task, we use the `show-snps` tool within MUMmer to call SNVs between all pairs of input genomes, which produces a distance matrix that can be visualized with PathogenDB-viz (see Results).

The same strategy is also used to rescale branches for our phylogenetic analysis, which we perform by creating a core genome alignment with `mugsy`,[47] a multiple genome aligner that internally combines MUMmer and its own algorithm for finding LCBs.[48] The core genome alignment in MAF format is converted to PHYLIP format with CLUSTALW,[49] and then this alignment undergoes maximum-likelihood phylogenetic inference via RAxML.[50] Since the outputted tree (in Newick format) initially has distances in units of time under the RAxML evolutionary model, which is more opaque than SNV distance, the tree's branches are rescaled by recalculating SNV distance using the aforementioned `show-snps` method on all adjoining nodes, including the ancestral states imputed by RAxML. Phylograms of these trees with overlaid SNV distances (as in Figure 2.3) can be plotted to PDFs using an included `mugsy_plot` task that wraps the `ape` R package.

[41] Darling, Mau, and Perna (2010).

[42] Tesler (2002), "GRIMM: genome rearrangements web server."

[43] Hannenhalli and Pevzner (1999), "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals".

[44] Lin and Moret (2008), "Estimating true evolutionary distances under the DCJ model".

[45] Hilker et al. (2012), "UniMoG-a unifying framework for genomic distance calculation: And sorting based on DCJ".

[46] Kurtz et al. (2004).

[47] Angiuoli and Salzberg (2011), "Mugsy: fast multiple alignment of closely related whole genomes."

[48] Recently, the Harvest suite was released for core genome alignment, which scales better to thousands of genomes than `mugsy`, and it even includes its own visualization tool, `gingr`. We are currently in the process of incorporating these tools into our pipeline. For more, see Treangen et al. (2014)

[49] Sievers et al. (2011), "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega".

[50] Stamatakis, Ludwig, and Meier (2005), "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees."

Although maximum likelihood phylogenetic analysis is a standard component of molecular epidemiology and typically the centerpiece of most published investigations of outbreaks using NGS,[51] it may in fact be overkill for answering the simpler day-to-day question of "are any new isolates closely related to the previously sequenced isolates?" By design, it requires a core genome alignment for all of the genomes that one wishes to include in the tree. Although every multiple sequence aligner uses heuristics to save time, multiple sequence alignment has a fundamental algorithmic complexity of $O(g^n)$ under the usual dynamic programming approaches,[52] where g is the average length of a genome and n is the number of genomes—i.e., exponential to the number of genomes. Even though tools for core genome alignment continue to get smarter and faster about subverting this complexity,[53] given the fundamental difficulties in scaling that problem, we anticipate that pairwise distance matrices will be a suitable alternative for outbreak detection as databases of thousands of assembled sequences become commonplace. Creating a distance matrix is guaranteed to be $O(n^2)$ in the most naive approach,[54] with each pairwise comparison being $O(g)$ by use of `show-snps`. Furthermore, adding one new assembled isolate does not require recalculating everything as in multiple sequence alignment, but we can instead add one row and column to the existing matrix in $O(2n)$ time. For these reasons we provide the `snv` task in conjunction with `mugsy`, and we rely on the distance matrices for analyses on >100 genomes, as presented later in the Results.

[51] Azarian et al. (2015), "Whole-genome sequencing for outbreak investigations of methicillin-resistant Staphylococcus aureus in the neonatal intensive care unit: time for routine practice?"; Eyre et al. (2012), "A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance."; Joensen et al. (2014); Casali et al. (2016), "Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study".

[52] Just (2001), "Computational complexity of multiple sequence alignment with SP-score."

[53] Treangen et al. (2014).

[54] And this is how it is currently implemented; precalculating the MLST for all isolates and only allowing within-MLST comparisons would be a simple first optimization.

PathogenDB-viz

Visualization is finally performed with the PathogenDB-viz toolkit. The interface, which has a heatmap layout and a geospatial layout, will be presented in the Results in Figures 4.7-4.10. PathogenDB-viz reads data from JSON files containing genetic distances and isolate metadata as prepared by PathogenDB-comparison's `heatmap` task, and displays it in a HTML5 interface that draws data dynamically to scalable vector graphics (SVG) using the d3.js Javascript library.[55] PathogenDB-viz is currently implemented as a single PHP page that loads all data via Asynchronous Javascript and XML (AJAX).[56] Given that all data is drawn on the client side, it attempts to maximize the control the user has over the selection of isolates to displayed and how the comparison is presented.

[55] `https://d3js.org/`

[56] Paulson (2005), "Web Applications with Ajax".

Agglomerative hierarchical clustering is performed in the browser using the `ml-hclust`[57] node.js package, using single linkage to emphasize the "chaining" of closely related genomes into clusters. The heatmap.js[58] library is used to draw density plots of epidemiological incidence in the geospatial layout. Sequenced isolates are displayed in the geospatial layout as a live-updating force-directed network using `d3.forceSimulation`, with a strong force keeping nodes from colliding, a moderate force pulling nodes toward the position of specimen collection, and a very weak spring force along the edges (which represent the putative transmissions under the selected SNV threshold).

[57] https://www.npmjs.com/package/ml-hclust

[58] https://www.patrick-wied.at/static/heatmapjs/

Availability

Source code for PathogenDB-pipeline, PathogenDB-comparison, and PathogenDB-viz is publicly available from GitHub at:

1. https://github.com/powerpak/pathogendb-pipeline
2. https://github.com/powerpak/pathogendb-comparison
3. https://github.com/powerpak/pathogendb-viz

The code in each repository is still under active development to suit the operational needs of the Pathogen Surveillance Program. The software is currently configured for execution on Mount Sinai's high performance computing environment (Minerva), but we will adapt it for simple installation on vanilla Linux distributions and provide machine images suitable for common cloud computing environments once we are ready to promote usage by other groups.

Results and Discussion

Assembly quality

As of April 2017, PathogenDB-pipeline has been used to assemble and annotate 593 genomes from 7 species. General statistics on assemblies produced by the pipeline are presented in Table 4.1. Most of the isolates assembled so far are *Staphylococcus aureus* and *Clostridium difficile* strains. About three quarters of the assembled *S. aureus* isolates were methicillin-resistant (MRSA). A few other rarer species have also been assembled.[59] 164 assemblies underwent manual curation during the development of the pipeline and are excluded from statistics on assembly quality.

[59] Note that the *S. maltophilia* isolates from Chapter 2 predate PathogenDB and therefore are not included in Table 4.1.

Characteristic	Assemblies (%), *N*=593
Species	
Clostridium difficile	221 (37.3)
Staphylococcus aureus	
Methicillin-resistant	262 (44.3)
Methicillin-resistant	90 (15.2)
Clostridium innocuum	4 (0.7)
Enterococcus faecium	4 (0.7)
Other	7 (1.1)
Assembly quality (uncurated assemblies only)	**Assemblies (%), *N*=429**
N50 > 1Mbp	412 (96.0)
Circular chromosome[a]	303 (70.6)
Largest contig size	
≥1Mbp	418 (97.4)
≥100kbp, <1Mbp	8 (1.9)
<100kbp	3 (0.7)
Number of contigs	
1	102 (23.8)
2	110 (25.6)
3	71 (16.6)
4	40 (9.3)
≥5	84 (19.6)

Table 4.1: **Statistics on assemblies generated by PathogenDB-pipeline since 2013.** Abbreviations: *N50*, shortest contig length above which 50% of the genome is included; Mbp, 1 million base pairs; kbp, 1 thousand base pairs.

[a] Any contig ≥1Mbp that circularized was considered a chromosome.

Without any curation (manual fixes), PathogenDB-pipeline is able to completely assemble most of the sequenced isolates, with >70% featuring a circular main chromosome. The *N50*, defined as the shortest contig at which it and all larger contigs would include 50% of the genome, was ≥1Mbp for 96.0% of uncurated genomes. 80.4% of completed assemblies contained four or fewer contigs, i.e., one chromosome (either closed or unclosed) and up to three plasmids or unassembled fragments. By these metrics, PathogenDB-pipeline is clearly able to produce many high-quality *de novo* assemblies without human intervention.

Computational benchmarks for PathogenDB-pipeline

Although over the past four years of development its end-to-end time has improved, PathogenDB-pipeline is still by far the most computationally intensive module within the PathogenDB suite. The majority of the cost is accrued during the *de novo* assembly and polishing steps, which require stepping through all read data for each sequenced isolate, which commonly exceeds 1Gbp per isolate. In Figure 4.6 we provide benchmarks for the impact of PathogenDB-pipeline on overall turnaround time for an end-to-end analysis and the corre-

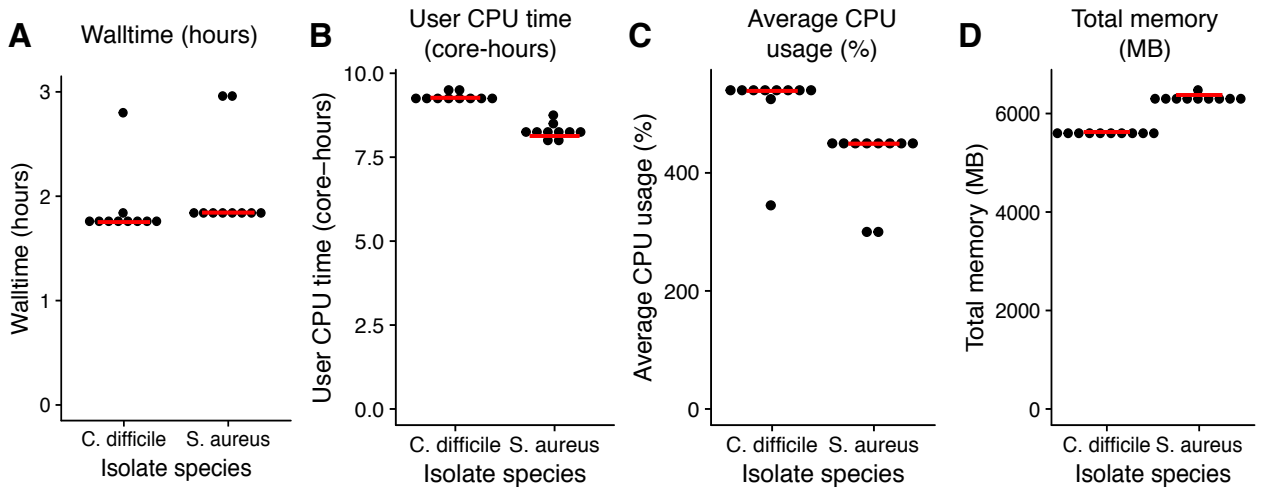


Figure 4.6: **Dotplots of computational benchmarks for PathogenDB-pipeline on a *S. aureus* and a *C. difficile* isolate.** For each isolate, measurements were collected from 10 end-to-end serial runs of the pipeline, starting at raw read data and ending at the `prokka_to_igb` task, using a single server with a 12-core Intel Xeon® 2.5GHz E5-2680 CPU and 128GB of RAM. Horizontal red lines indicate the median.

sponding computational cost. We performed 10 serial end-to-end runs of the pipeline starting from raw read data for one *S. aureus* isolate and one *C. difficile* isolate on a 12-core Intel Xeon® 2.5GHz E5-2680 server with 128GB of RAM. Most of the runs produced nearly identical benchmarks. The median walltime (which measures real-world start to end time) was under two hours, with none of the runs exceeding three hours (Figure 4.6A). The jobs benefited from multicore usage, as the median user CPU time in core-hours exceeded the walltime by a factor of 4-5× (Figure 4.6B), and this is confirmed by checking average CPU usage, which is normalized against a single core and stayed mostly in the 400-600% range (Figure 4.6C). Although HGAP, BLASR, and Quiver are memory-intensive steps, the total memory used did not exceed 7GB for any of the runs (Figure 4.6D).

PathogenDB-viz characterizes local outbreaks of S. aureus

After comparing a large number of same-species isolates with PathogenDB-comparison, the analyses can be presented to clinicians in an interactive visualization using PathogenDB-viz. Figure 4.7 displays the interface of PathogenDB-viz for a heatmap visualization of all *S. aureus* isolates that clustered with at least one other patient's isolate(s) at a SNV threshold of ≤10 SNVs. The software provides a web interface with many controls for quickly "drilling down" to the time range and isolates of interest. We now briefly tour this interface.

At top left, the user can select from analyses that were generated by the

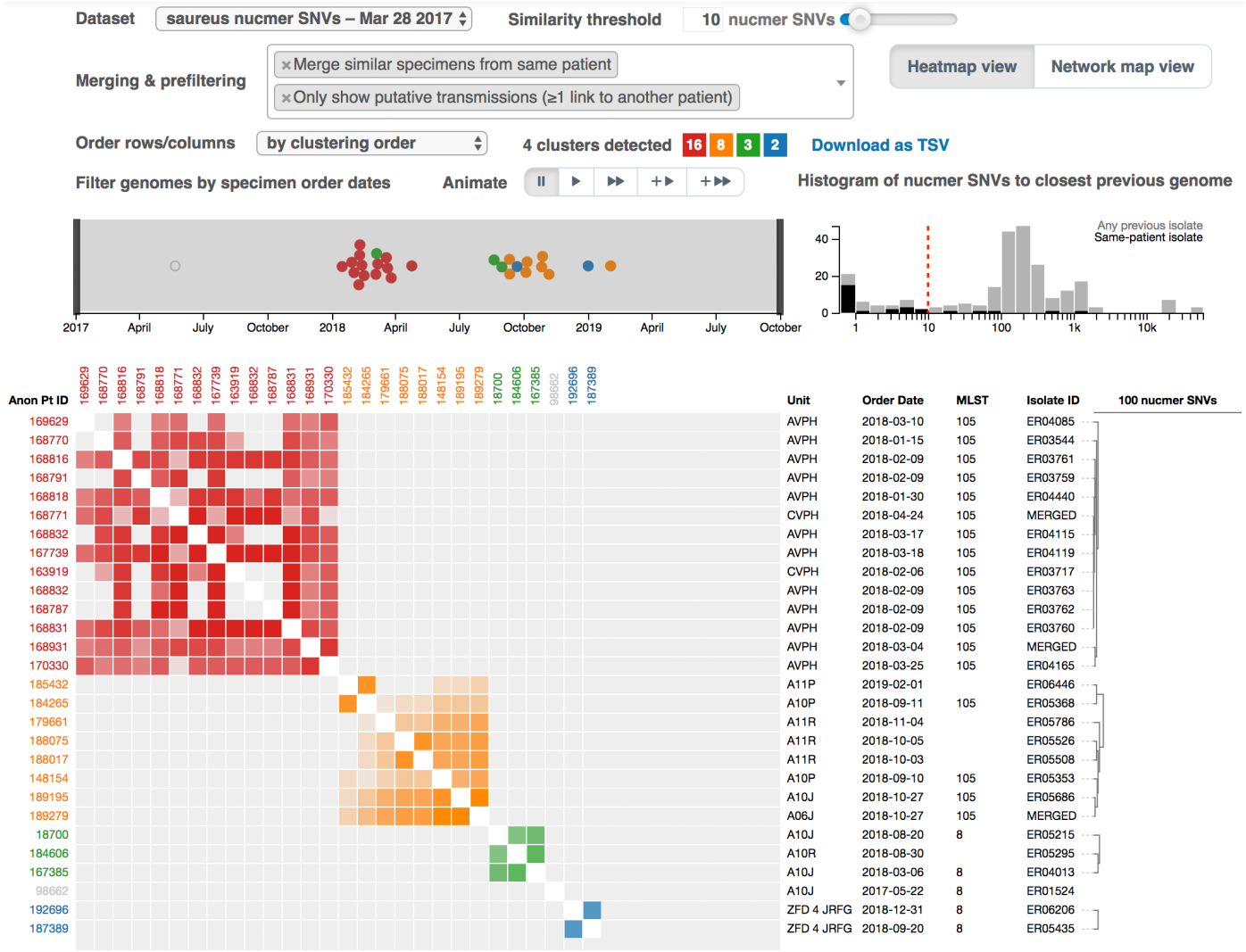


Figure 4.7: **PathogenDB-viz heatmap visualization for all putatively transmitted *S. aureus* isolates, based on NGS.** Note that dates have been shifted into the future and unit names have been obfuscated to reduce the disclosure of potentially identifying information. At top, user controls allow selection of the dataset, SNV threshold, merging and prefiltering of isolates, and ordering of the diagram. A horizontal beeswarm shows the distribution of isolates over collection times, which the user can "brush" to select specific time ranges. At adjacent right, a histogram of SNV distances between each genome and its closest previous neighbor helps inform what a reasonable SNV threshold might be; in black, isolates from the same patient (which are expected to be related); in gray, isolates from any previous patient. At bottom left, a clustered heatmap depicts distances between isolates that exceed the SNV threshold as the large colored blocks along the diagonal. At bottom right, a hierarchical clustering shows SNV distances between rows in the heatmap.

`heatmap` task of PathogenDB-comparison. The top right has a slider used to set a threshold in SNVs for considering two isolates to be related enough for putative transmission. While previous studies suggest using a very low threshold, e.g., 2-3 SNVs for *C. difficile* isolates collected within one year,[60] these studies used short-read sequencing and alignment-based methods, which miss SNVs in regions that can't align to the chosen reference—particularly structural variants, plasmids, and long repeats.

[60] Eyre et al. (2013), "Diverse sources of C. difficile infection identified on whole-genome sequencing."; Price et al. (2014), "Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of Staphylococcus aureus in an intensive care unit."; Eyre et al. (2012).

To justify a selected SNV threshold for a particular dataset, we provide a built-in analysis similar to what is presented in recent studies[61] as the histogram in the top right of the interface. This histogram shows the distribution of SNV distances from every genome to its closest (by SNV distance) neighbor among chronologically previous isolates, comparing the distributions for same-patient isolates (black) and different-patient isolates (gray). Actual transmissions should be reflected as a gray peak toward the left of the diagram (small distances), while the natural diversity of *S. aureus* in the community creates a separate peak toward the center. Indeed, in our data, we see a clear bimodal distribution (Figure 4.7). Also, as same-patient isolates are expected to be share lineage, the left-side black peak serves as the "positive control" for distances that are representative of clonality in our data (and therefore would also imply transmission for different patient isolates). Based on the overlapping left-side peaks and the midpoint between the two gray peaks, a cutoff of 10 SNVs appears reasonable for the depicted 2.7 year period to define continuity of lineage. This is consistent with the 5-10 SNV per year average mutation rate observed in recent NGS surveys of hospital-associated *S. aureus*.[62]

[61] See Figure 1A of Eyre et al. (2013).

[62] Price et al. (2014); Harris et al. (2013).

The user can specify merging and prefiltering options for isolates at the top left. The "merge similar specimens from the same patient" option is particularly useful for condensing the display, as it collapses all same-patient isolates under the SNV threshold into single datapoints.[63] This allows the user to focus on only the links between different-patient isolates. Similarly, the option to "only show putative transmissions" hides all isolates with no links to a different-patient isolate under the SNV threshold. What remains, in effect, are only the sequenced patient isolates involved in putative transmission events.

[63] Resampling is common in our dataset since some patients get standing orders for daily cultures and the Pathogen Surveillance Program receives all positive *S. aureus* culture specimens.

PathogenDB-viz automatically performs clustering of the isolates that match the criteria. In Figure 4.7, over the 2.7 year period we see that four clusters were detected, with sizes of 16, 8, 3, and 2 patients. One of these clusters (red) was

already known to infection prevention and control staff and resulted in the shutdown and deep cleaning of the involved hospital unit. The timeline beeswarm plot shows that all of these isolates fell within a roughly four month span, and the end of this span is when the cleaning occurred; thankfully, no new isolates related to this cluster have been detected since. The 8 patient cluster (orange), which spans a more recent interval of six months, was not known to infection prevention staff before sequencing occured, and in fact was discovered by use of this visualization.

The main area of the visualization (bottom left) shows a clustered heatmap of distances between the isolates that matched the filtering criteria. Each patient is both a row and a column, and a link between two patients underneath the SNV threshold results in a colored box. (Although not shown in Figure 4.7, these boxes can be clicked to reveal more detailed information about each patient isolate with links to corresponding records in PathogenDB.) Almost all isolates within a cluster should be related to each other underneath the SNV threshold, which results in large colored squares along the diagonal. The presence and size of these squares are an easy way for infection prevention and control officers viewing the data to judge how many "clusters" of transmission appear to be present in the time interval and the level of evidence for the coherence of a cluster. Pairwise MUMmer comparisons can result in spurious SNV calls in certain hypervariable regions (e.g., phage elements), which artificially inflates SNV distances and results in "holes" in the square (as in the red cluster).

The hierarchical clustering distances are shown as a more familiar tree at the bottom right of the visualization, along with metadata for each isolate like order date, MLST, and unit of collection. In this case, the metadata reveal that the red cluster was concentrated in a single unit (which is how infection prevention became aware of it via epidemiological data alone). The 8-patient orange cluster, however, is spread across multiple units.

A SPATIAL LAYOUT can be useful to visualize isolates and their links in relationship to hospital locations, and is provided as an alternative view by PathogenDB-viz (Figure 4.8). This interface, which is accessed by toggling the "Network map view" button in the upper right corner, has been focused on only the isolates in the time range of the orange cluster, although we now include unrelated isolates as well (note that the beeswarm timeline contains new light gray

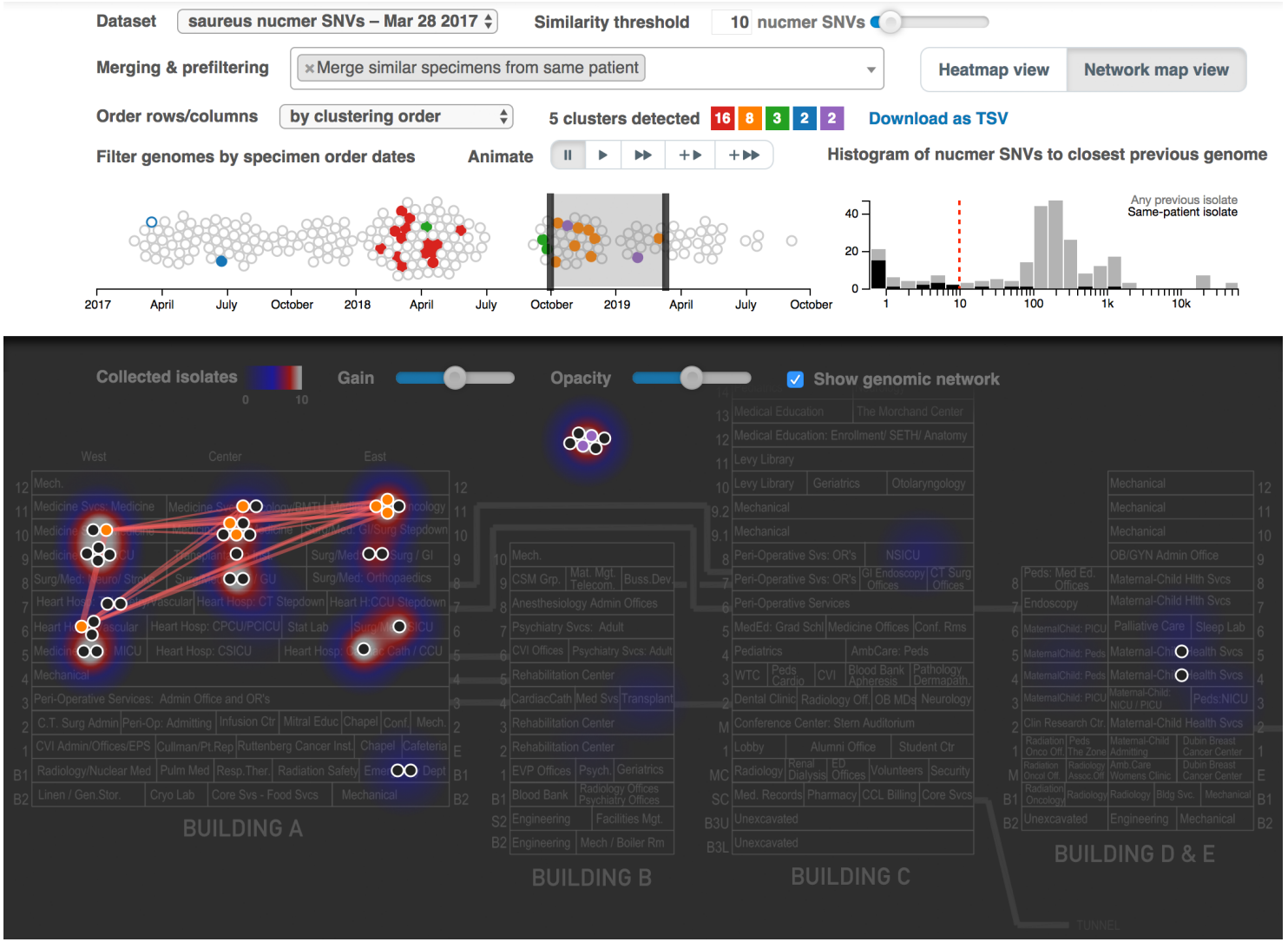


Figure 4.8: **PathogenDB-viz geospatial visualization for an NGS-confirmed cluster of *S. aureus* isolates.** Note that dates have been shifted into the future to reduce the disclosure of potentially identifying information. The top of the interface is as in Figure 4.7. At bottom, node-link diagram of putative transmissions, with circles representing sequenced isolates in the selected time range (see timeline beeswarm plot) and red lines connecting isolates underneath the chosen SNV threshold (here, 10 SNVs). Nodes are placed on top of the hospital unit (light gray stacking diagram, vertical axis corresponds to floor level) from which the isolate was collected. The density plot underneath nodes represents the overall frequency of positive cultures at each location of the hospital.

circles, which are the isolates that were not involved in any putative transmissions). In this view, the isolates are depicted as dots on top of a stacking layout of Mount Sinai's hospital campus, which has several different buildings (see the Building labels at bottom; names anonymized). The floors are laid out vertically in this diagram, and connections between them (like sky bridges) are depicted as thick lines. (The isolates hovering over Building B were sent from a different campus.) Dots are colored by the cluster they were in, and transmissions between patients according to the SNV threshold are drawn as red lines. This diagram makes it apparent that the patients in the orange cluster were spread across the upper floors of the leftmost building, with at most three patients sharing a unit. The spatial network layout also includes a density plot (the fuzzy clouds underneath the points) that depicts all collected isolates in PathogenDB, including those not yet sequenced. Therefore, the density plot can reveal parts of the hospital that had many cases of the HAI, but have been undersampled by the NGS data, and could merit follow-up sequencing of the banked isolates. In the case of Figure 4.8, there are no clouds without dots, meaning that all "hot spots" for *S. aureus* in the hospital during this time interval were at least partially captured in the NGS data.

PathogenDB-viz identifies local diversity and transmissions of C. difficile

We now use PathogenDB-viz to similarly visualize all sequenced isolates of *C. difficile*. While the Pathogen Surveillance Program has collected all positive *S. aureus* cultures over the past ~2 years, the surveillance of *C. difficile* has been broader, including sequencing of isolates from pilot periods roughly one year and three years before the start of routine daily collection from all Mount Sinai units ~1.5 years ago. This is reflected in the trimodal distribution of isolates along the timeline in Figure 4.9, where we have included all isolates, not just putative transmissions. We set a similar SNV threshold of 10 SNVs for this longer time period based on the histogram of SNV distances (top right of Figure 4.9). Under this threshold, we discover ten small clusters, the largest of which has four patients and with most having only two. Scanning the color of the dots in the timeline indicates that most of the clusters sensibly fall within month-scale intervals, although the yellow-green, gray, and blue clusters are remarkable for spanning the >1 year gaps between the different surveillance periods.

Figure 4.9: **PathogenDB-viz heatmap visualization for all sequenced *C. difficile* isolates over a five-year period.** All conventions used here are equivalent to Figure 4.7. Note that dates have again been shifted into the future to reduce the disclosure of potentially identifying information, and we have also censored all sample metadata from this screenshot.

The heatmap visualization at the bottom left of Figure 4.9 shows the clusters as small groups along the diagonal, and the clustering dendrogram at bottom right shows that some of the clusters separate from each other by distances of ~20 SNVs.[64] Therefore, there is certainly some interplay between *C. difficile* evolution in the local community and what eventually arrives at Mount Sinai. What is more notable is that many isolates across the five year capturing period appear to be unrelated (the heatmap is mostly blank). Although constrained by our limited sampling, this is consistent so far with a previous NGS study that concluded that patient-to-patient transmissions were not a major source of hospital-associated *C. difficile* infections over a three-year period in Oxfordshire, United Kingdom.[65]

[64] Recall the current estimates of the mutation rate for *C. difficile* are around 2-3 SNVs per year; see Eyre et al. (2013) and Eyre et al. (2012).

[65] Eyre et al. (2013).

Note that there are three spurious low SNV distances seen in the upper left of the heatmap (three unclustered black squares) likely caused by a low-quality assembly, which should be re-examined. Low-quality assemblies can appear to have no SNVs if they contain too much redundant sequence, usually caused by duplicate contigs failing to be merged during assembly. Since the heatmap calculates SNV distances in both directions (above and below the diagonal), the lack of an equivalent SNV distance in the reverse comparison provides an easy way to double-check the procedure. This offers some resilience against spurious low SNV transmission predictions caused by misassembled genomes.

Finally, we can again visualize the spatial relationships among clusters using the network map view in Figure 4.10. In this view, which is now focused only on isolates with at least one link to a different-patient isolate, we see that most of the clusters are spread across multiple units (red lines). There are even NGS-confirmed links between units in different buildings. Although considering all the hypothetical reasons for this pattern is beyond the scope of this chapter, if these transmissions of *C. difficile* spores are occurring with Mount Sinai and not within the community, they are taking places across substantial spatial distances, whether due to movements of patients, visitors, staff, or equipment.

Conclusions

We have developed a new open-source software suite, PathogenDB, that permits semi-automated epidemiological analysis of HAIs based on long-read sequencing and *de novo* assembly of all isolates. We chose a modular design, sep-

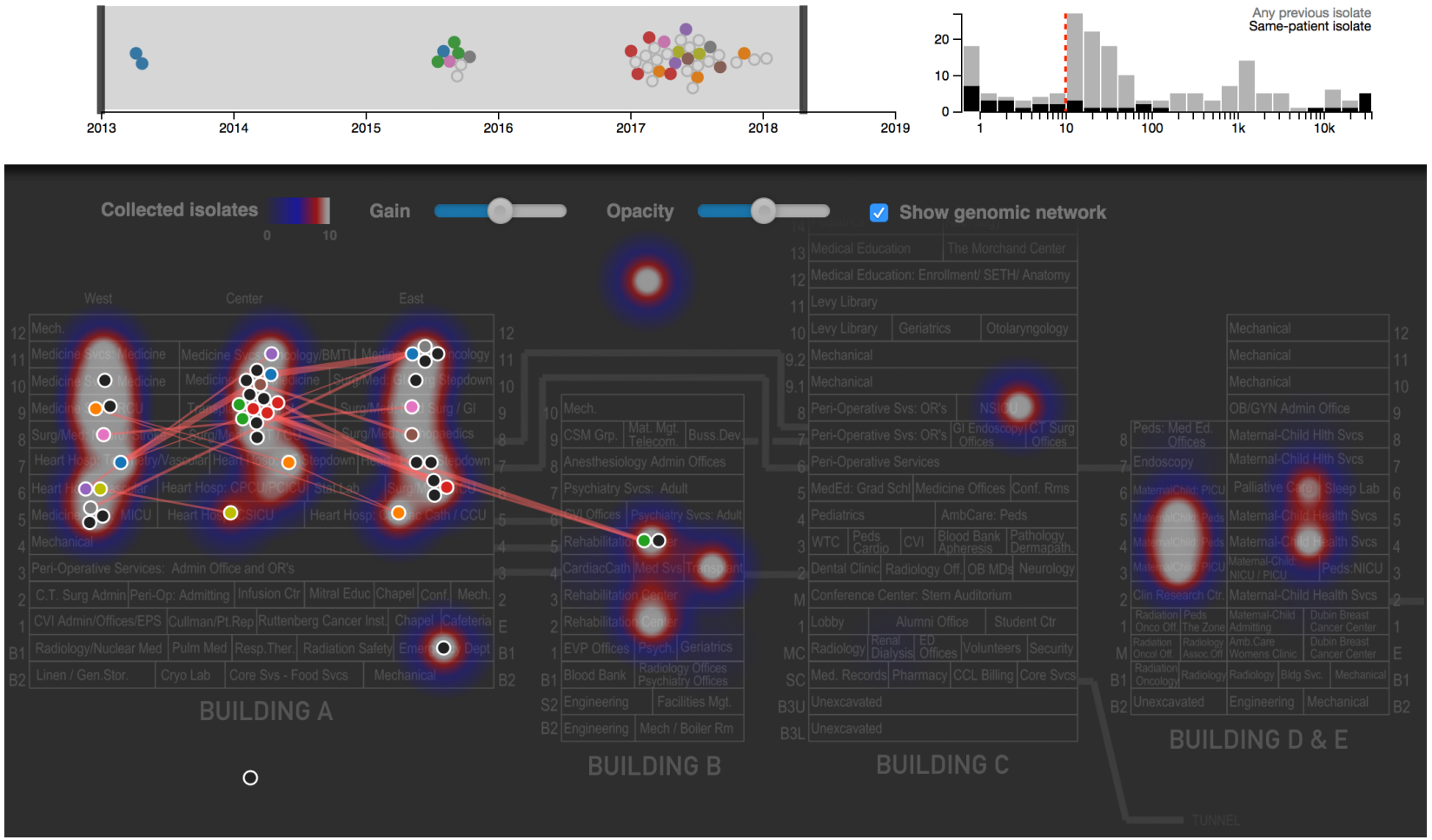


Figure 4.10: **PathogenDB-viz geospatial visualization for NGS-confirmed clusters of *C. difficile* isolates over a five-year period.** All conventions used here are equivalent to Figure 4.8. Note that dates have been shifted into the future to reduce the disclosure of potentially identifying information.

arating concerns into a LIMS that centralizes storage of the most up-to-date data, a genome assembly and annotation workflow called PathogenDB-pipeline, a comparative genomics toolkit called PathogenDB-comparison, and the visualization tool PathogenDB-viz. Thus far,we have assembled and annotated 593 genomes, mostly of *S. aureus* and *C. difficile*, using PathogenDB-pipeline. This part of the analysis, which is the most computationally intensive, can run in 2-3 hours per isolate on a single server (Figure 4.6) and in >70% of cases can completely finish the genome assembly without human intervention (Table 4.1). We can then integrate phylogenetic and epidemiological analyses into a "live view" of putative transmissions mapped to hospital locations, using novel interactive heatmap and network map layout visualizations as implemented in PathogenDB-viz.

Our software suite was able to genomically characterize one known MRSA outbreak (red cluster in Figure 4.7) and discover a previously unknown MRSA outbreak (orange cluster in Figures 4.7-4.8). We have likewise used it to characterize the incidence and spatial distribution of transmissions of *C. difficile* across a surveillance period of five years (Figures 4.9-4.10). Addtionally, we used earlier versions of our software to characterize two transmissions via solid organ transplant,[66] and pseudo-outbreaks of *B. cepacia* and *S. maltophilia*.[67] The three modules of the PathogenDB suite are freely available on GitHub (see Availability above) and are being prepared for public use in generic computing environments.

[66] Altman et al. (2014), "Transmission of Methicillin-Resistant Staphylococcus aureus via Deceased Donor Liver Transplantation Confirmed by Whole Genome Sequencing."; Bashir et al. (2017), "Genomic confirmation of vancomycin-resistant Enterococcus transmission from deceased donor to liver transplant recipient."

[67] Pak et al. (2015), "Whole-Genome Sequencing Identifies Emergence of a Quinolone Resistance Mutation in a Case of Stenotrophomonas maltophilia Bacteremia."

Notes

Contributions

Theodore R. Pak (TRP), Mitchell Sullivan (MS), Oliver Attie (OA), Elizabeth Webster (EW), Robert Sebra (RS), Camille L. Hamula (CLH), Gintaras Deikus, (GD), Leah C. Newman (LCN), Gopi Patel (GP), Deena R. Altman (DRA), Shirish Huprikar (SH), Ali Bashir (AB), Andrew Kasarskis (AK), and Harm van Bakel (HVB) contributed to this chapter.

TRP wrote the first versions of PathogenDB-pipeline, PathogenDB-comparison, and PathogenDB-viz. TRP, MS, OA, HVB, AB, and EW have contributed code to the current version of PathogenDB-pipeline. TRP, MS, OA, HVB, and EW have contributed code to the current version of PathogenDB-comparison.

TRP wrote the current version of PathogenDB-viz. HVB created and maintains the PathogenDB database. DRA, GP, HVB, AB, and CLH organized collection of samples. GD, LCN, and RS prepared sequencing libraries and performed sequencing. CLH performed culturing and drug susceptibility testing. TRP, MS, OA, RS, EW, HVB, and AB performed data analysis. SH, GP, and AK provided institutional support and critical feedback on PathogenDB-viz. TRP created all figures and wrote the first draft of this chapter.

Funding

Funding was provided by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai. TRP was supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai and NIH/NIAID (U19-AI118610 and F30-AI122673). Research was also supported by the Office of Research Infrastructure of the NIH under award number S10-OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

We thank Timothy O'Donnell, Tavi Nathanson, and members of the clinical microbiology laboratory at Mount Sinai for their contributions. This work was supported in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

5

Estimating local costs of Clostridium difficile *infection using statistical learning and electronic medical records*

Reported per-patient costs of Clostridium difficile *infection (CDI) vary by two orders of magnitude among different hospitals, suggesting that precise, local analyses are needed to guide decision-making. We sought to estimate changes in length of stay (LOS) associated with CDI at one hospital using only automatically extractable electronic medical record (EMR) data and performed a retrospective cohort study of 171,938 visit records spanning a 7-year period. 23,968 variables were extracted from EMR data recorded within 24 hours of each admission to train an elastic net regularized logistic regression model for propensity score matching. To address time-dependent bias (reverse causation), we stratified comparisons by time-of-infection and fit multistate models. The estimated difference in median LOS for propensity-matched cohorts varied from 3.1 days (95% CI, 2.2–3.9) to 10.1 days (95% CI, 7.3–12.2) depending on the case definition; however, dependency of the estimate on time-to-infection was observed. Stratification by time to first positive toxin assay, excluding probable community-acquired infections, showed a minimum excess LOS of 3.1 days (95% CI, 1.7–4.4). Under the same case definition, the multistate model averaged an excess LOS of 3.3 days (95% CI, 2.6–4.0). Changes in LOS can be extrapolated to a marginal dollar cost per CDI case by multiplying by the average cost of an inpatient-day. We conclude that infection control officers can leverage automatically extractable EMR data to estimate costs of CDI at their specific institution.*

Mr. Marks, by mandate of the District of Columbia Precrime Division, I'm placing you under arrest for the future murder of Sarah Marks and Donald Dubin that was to take place today, April 22 at 0800 hours and four minutes.

—JOHN ANDERTON, *Minority Report*

C-3PO: *Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to one.*

HAN SOLO: *Never tell me the odds!*

—Star Wars: Ep. V – The Empire Strikes Back

Clostridium difficile INFECTION (CDI) is the most frequently reported healthcare-associated infection (HAI) in the US[1] and the major infective cause of nosocomial diarrhea in developed countries,[2] incurring billions of dollars in excess medical costs per year.[3] Estimates of the per-patient cost of CDI have varied from $2,871 to $122,318 due to differences in methodology, patient in-

[1] Leffler and Lamont (2015), "Clostridium difficile."

[2] Davies et al. (2014), "Underdiagnosis of Clostridium difficile across Europe: The European, multicentre, prospective, biannual, point-prevalence study of Clostridium difficile infection in hospitalised patients with diarrhoea (EUCLID)"

[3] Zimlichman et al. (2013), "Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system."

clusion criteria, and regional costs.[4] Given the high hospital-to-hospital variability of these costs,[5] infection control officers, hospital administrators, and clinicians would benefit from estimates tailored to their particular population and healthcare practices. Concretely defining the potential economic savings of CDI prevention would empower stakeholders to prudently choose among the many available validated interventions.[6]

Measuring costs within healthcare systems is notoriously difficult (particularly in the US), as many hospitals do not have access to structured, itemized reimbursement data linked to all of their patient medical records.[7] Even the institutions that have informatics capabilities to retrospectively link these data have relied on the curation of select variables and chart review to estimate attributable CDI cost.[8] Nevertheless, electronic medical record (EMR) systems are used by the majority of first-world acute care facilities.[9] Part of the rationale for these systems is that hospitals may leverage EMR data for optimal decision-making by inferring causal relationships from raw observations during routine care.[10] An analysis based on automatically extractable data from an EMR that quantifies preventable hospital costs, such as those attributable to an HAI like CDI, would be of great value in building a continuously learning healthcare system.[11] EMRs contain many structured fields relevant to this analysis, including: diagnosis codes and lab results demonstrating onset of HAIs; thousands of variables for procedures, problems, and medications that can serve as covariates for adjustment in observational studies; and importantly, the length of stay (LOS) for each visit, which is the primary contributor to excess costs for most HAIs, including CDI.[12]

The goal of this study was to generate a robust estimate of local cost associated with CDI using data that are automatically extractable from a typical EMR. We use all available structured data recorded within 24 hours of admission in the EMR—including over 20,000 variables, such as medications reported and administered, abnormal lab values, and problem list entries—to build fully data-driven models for CDI risk using a machine learning algorithm, avoiding the potential bias of preselected covariates and manual chart review. CDI risk models trained on uncurated data from EMRs have already outperformed models that only incorporate variables for known risk factors, indicating that CDI risk may be nuanced in particular care settings.[13] We then use these trained CDI risk models for propensity score matching, which allows estimation of changes

[4] Gabriel and Beriot-Mathiot (2014), "Hospitalization stay and costs attributable to Clostridium difficile infection: A critical review"; Ghantoji et al. (2010), "Economic healthcare costs of Clostridium difficile infection: A systematic review"; Zhang et al. (2016), "Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study".

[5] Lofgren et al. (2014), "Hospital-Acquired Clostridium difficile Infections"; Stevens et al. (2015), "Excess Length of Stay Attributable to Clostridium difficile Infection (CDI) in the Acute Care Setting: A Multistate Model."

[6] Dubberke et al. (2014), "Strategies to Prevent Clostridium difficile Infections in Acute Care Hospitals: 2014 Update"; Katz (2013), "Pay for preventing (not causing) health care-associated infections."

[7] Cooper et al. (2015), "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured".

[8] Dubberke et al. (2008), "Short- and Long-Term Attributable Costs of Clostridium difficile-Associated Disease in Nonsurgical Inpatients"; Dubberke et al. (2014), "Attributable inpatient costs of recurrent Clostridium difficile infections"; Greco et al. (2015), "Costs associated with health care-associated infections in cardiac surgery".

[9] Gray et al. (2011), "Electronic health records: an international perspective on "meaningful use"."; Henry et al. (2016), "Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015".

[10] Dahabreh and Kent (2014), "Can the learning health care system be educated with observational data?"; Etheredge (2007), "A rapid-learning health system."; Pak and Kasarskis (2015), "How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management".

[11] Krumholz, Terry, and Waldstreicher (2016), "Data Acquisition, Curation, and Use for a Continuously Learning Health System".

[12] McGlone et al. (2012), "The economic burden of Clostridium difficile"; Wilcox et al. (1996), "Financial burden of hospital-acquired Clostridium difficile infection."; Zimlichman et al. (2013).

[13] Wiens, Guttag, and Horvitz (2014), "A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions."

in LOS associated with CDI. Most previous studies of CDI cost do not account for the possibility that longer LOS increases the risk of CDI, i.e., reverse causation, and therefore likely overestimate the cost of CDI.[14] To adjust for this, we stratify our analysis by the time of CDI diagnosis to find the change in LOS conditional on minimal prior exposure to the hospital environment. Finally, we compare these results to a multistate model of competing time-dependent risks between discharge and the onset of CDI.

[14] Mitchell et al. (2014), "The prolongation of length of stay because of Clostridium difficile infection"; Stevens et al. (2015).

Methods

Data source

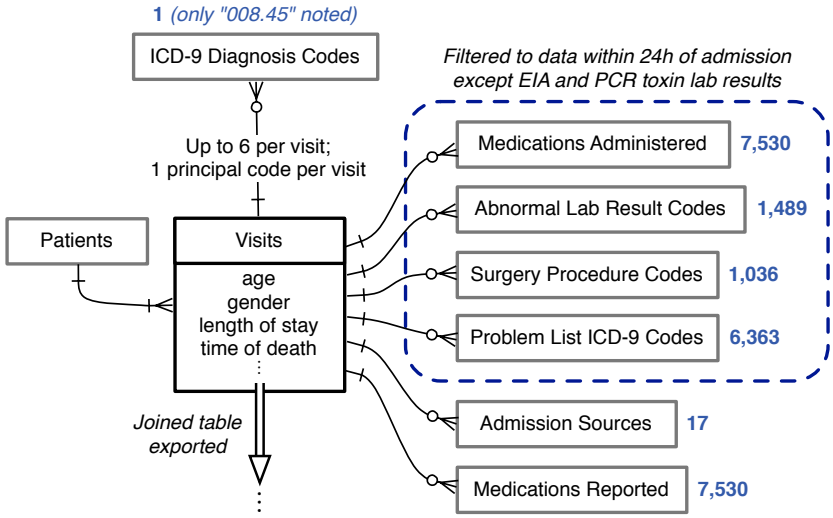


Figure 5.1: **Data sources for this study**. Entity-relationship diagram for all electronic medical record data used to generate models of *Clostridium difficile* infection propensity, using Information Engineering notation; see Halpin and Morgan (2010). Boxes represent tables of entities with any directly associated attributes (fields) listed below; single lines represent relationships, with arrowheads indicating the cardinality of each side of the relationship; crow's foot arrowhead with circle represents "zero or more;" crow's foot arrowhead with cross-stroke represents "one or more;" cross-stroke arrowhead represents "exactly one." Blue numbers indicate the number of variables extracted from each associated table for each visit. ICD-9, International Classification of Diseases Ninth Revision; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

This study was conducted at The Mount Sinai Hospital, a 1,171-bed tertiary care hospital in New York, NY. Records of adult inpatient visits were extracted from warehoused Epic EMR data and de-identified using the HIPAA Safe Harbor method, 45 CFR §164.514(b)(2). Data was collected on demographics, LOS, time of death, admission sources, reported medications, and the presence of a "008.45" ICD-9 principal or secondary visit diagnosis code denoting "Intestinal infection due to *Clostridium difficile*." Furthermore, all records of medications administered, abnormal lab result codes, surgery procedure codes, or problem list ICD-9 codes within the first 24 hours after admission were collected as boolean variables (presence or absence). All codes and variables that were uni-

form across the study population were dropped from the dataset. The relationship between collected data elements, including the maximal cardinality of each datatype, are summarized in Figure 5.1. This study was approved by Mount Sinai's Institutional Review Board as exempt research.

Study population

The cohort included all patients 18 years of age or older admitted between January 1, 2009 and October 22, 2015. For each patient, visits following the first recorded visit in the time range were excluded so that each patient corresponded to a single visit. Visits involving a patient death, defined as a recorded time of death within 24 hours after discharge, were excluded. Visits with missing or invalid date information were excluded (<0.01% of all records).

Study design

Prior studies vary on the use of ICD-9 discharge codes vs. positive laboratory tests to define CDI cases[15] and identify differing positive predictive values for immunoassay and nucleic acid based laboratory tests.[16] To ensure maximally robust results and allow comparison with prior studies, we repeated our analysis for five definitions of CDI:

(i) An "008.45" principal or secondary ICD-9 visit diagnosis code
(ii) ≥1 positive stool toxin enzyme immunoassay (EIA) lab result
(iii) ≥1 positive stool toxin polymerase chain reaction (PCR) lab result
(iv) Either ii or iii
(v) Any of i, ii, or iii

Our study's time range included both a period where the EIA assay was the standard hospital laboratory test (~3 years) followed by a period where the PCR assay was standard (~4 years). For case cohorts (ii) and (iii), comparisons were only permitted with controls from the time range during which that same test was standard. The hospital laboratory only performs toxin assays on unformed stool samples, implying the presence of diarrhea for positive results.

Statistical analysis

Propensity models for CDI based on the five case definitions were trained using logistic regression with elastic net regularization. For model coefficients β and

[15] Gabriel and Beriot-Mathiot (2014); Zhang et al. (2016).

[16] Bagdasarian, Rao, and Malani (2015), "Diagnosis and Treatment of Clostridium difficile in Adults"; Moehring, Lofgren, and Anderson (2013), "Impact of Change to Molecular Testing for Clostridium difficile Infection on Healthcare Facility-Associated Incidence Rates."; Polage et al. (2015), "Overdiagnosis of Clostridium difficile Infection in the Molecular Test Era."

N observations of ρ predictors x and the output variable y, elastic net regularization aims to solve

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right]$$

over a grid of values of λ, which controls the overall penalization for model size.[17] Here $l(y, \eta)$ is the negative binomial log-likelihood contribution for observation i, which is defined as

$$y_i \cdot (\beta_0 + x_i^T \beta) - \log(1 + e^{(\beta_0 + x_i^T \beta)}).$$

The α hyper-parameter, controlling the ratio of ℓ_1 to ℓ_2 penalties, was empirically selected during model fitting for the first case definition using a grid search, maximizing mean area under the receiver operating characteristic curve (AUROC) under five-fold nested cross validation (all values for $\alpha \leq 0.1$ produced equivalent AUROCs) and checking for effective shrinkage of coefficients (to within one-fifth of the size of the simplest model). This resulted in selecting $\alpha = 0.03$. The λ hyper-parameter was empirically selected for each modeled case definition by maximizing mean AUROC under five-fold cross validation (Figure 5.2). Nested cross validation while selecting λ was used to evaluate AUROC (which are the performance values reported in Results), while the final propensity models were allowed to see the entire training dataset. 100 bootstrap resampling runs were used to estimate the 95% AUROC confidence interval (CI). Since propensity models were intended to fairly assess risk at admission for CDI across both cases and controls, variables that could reflect workup or treatment of CDI (as opposed to pre-existing risk) were masked (Table A.1). A board-certified infectious diseases physician reviewed the final set of variables selected by each model after regularization[18] to ensure that none of them reflected medical sequelae of CDI as opposed to potential risk factors.

Matching (1:1) on the propensity score was performed without replacement of controls using a nearest neighbor-matching algorithm and a caliper of 0.2 standard deviations of the logit of the propensity score,[19] after exact matching on gender and age divided into six ranges. To assess the performance of the matching, we calculated standardized mean differences[20] for age and gender, for which a difference between -0.1 and 0.1 is generally considered negligible,[21] and examined the distributions of the propensity scores between matched groups.

[17] This penalization alleviates the degeneracies of logistic regression when the number of predictor variables $p > N$.

[18] See Supplementary Data S1.

[19] Austin (2011), "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies."

[20] Ibid.

[21] Haukoos and Lewis (2015), "The Propensity Score".

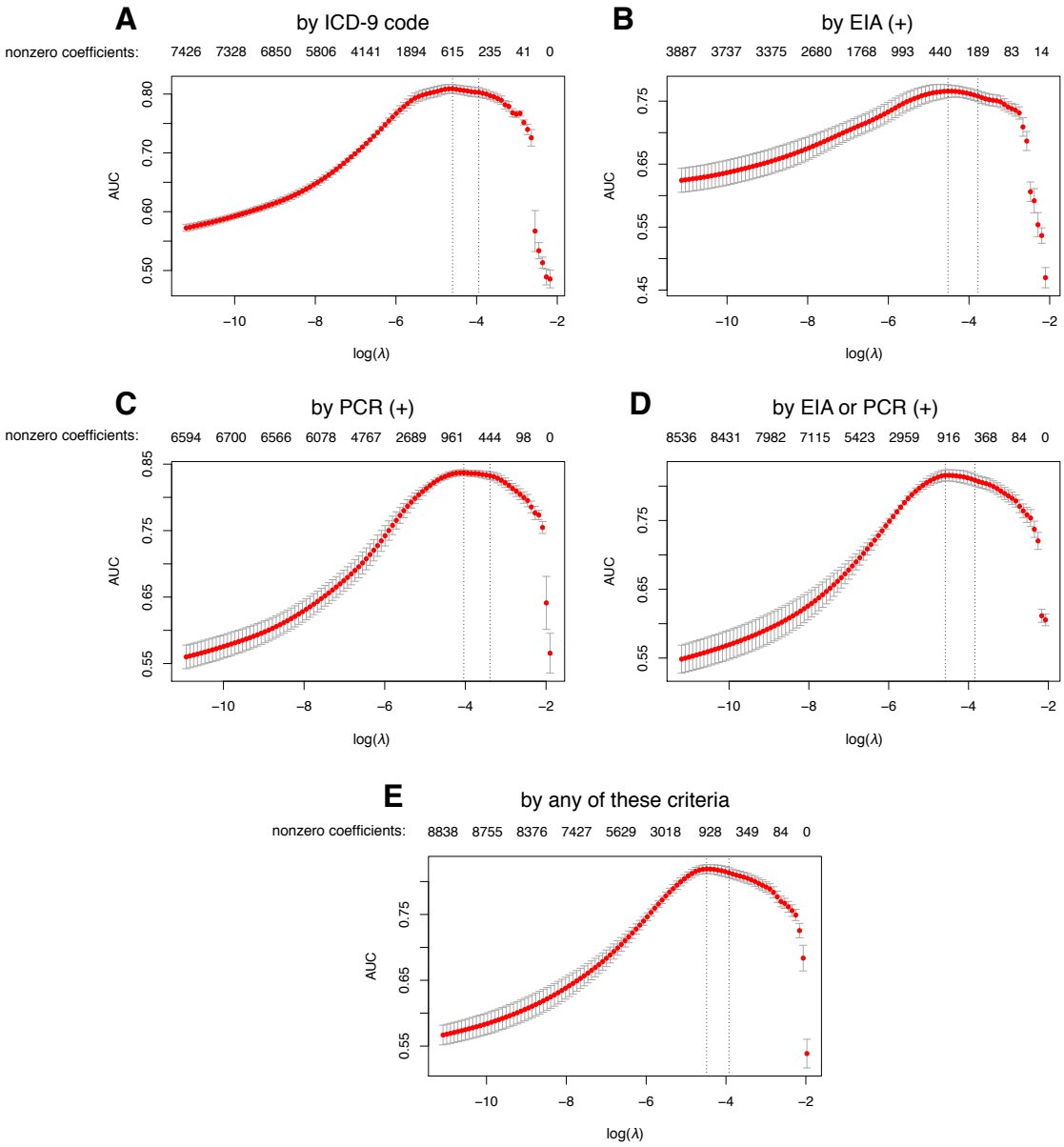


Figure 5.2: **Empirical selection of the regularization penalty hyper-parameter λ.** A–E, mean area under the receiver operating characteristic curve (AUC) vs. $\log(\lambda)$ for each of the case definitions for *C. difficile* infection, based on five-fold cross-validation. As the $\log(\lambda)$ penalty hyper-parameter increases, the number of variables left in the model (i.e., with nonzero coefficients) decreases; these are listed across the top of each plot. Vertical bars indicate standard errors. The $\log(\lambda)$ value with maximal mean AUC and the $\log(\lambda)$ value with mean AUC within 1 standard deviation of the maximum are highlighted by vertical dashed lines; the former was used for the final propensity models. ICD-9, International Classification of Diseases Ninth Revision; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

Furthermore, to ensure that propensity matching itself does not cause spurious changes in the outcome variable (LOS), we repeated the matching algorithm using the matched controls against remaining unmatched controls, creating a "matched-again" control cohort, with the expectation that re-matching controls should not, by itself, create significant differences in the outcome variable. For each case definition of CDI, differences of the median length of stay (LOS) between cases and matched controls were calculated and 95% CIs were calculated from 10,000 bootstrap resampling runs. The minimum LOS was specified as 0.1 days (2.4 hours), with smaller values rounded up to this value. Statistical significance of differences in LOS was determined using two-sided Mann-Whitney *U* tests. *P* values for all contrasts reported in Figures 2 and 3 are conservatively Bonferroni-corrected for the full number of hypotheses (24). Kaplan-Meier plots were generated to examine the nonparametric maximum likelihood estimation for risk of discharge from the hospital between cases and matched controls, and 95% CIs for these plots were derived from log-transformed standard errors.

To examine whether changes in LOS may have depended on the time of CDI onset, we repeated the above analysis for case definition (iv) stratified by the time of the first positive toxin assay result, using three ranges: 0–3 days, 3–8 days, and ≥8 days. Propensity models were again fitted to each of these case cohorts for matching as described previously, with the added condition that controls discharged before the start of the CDI time window were ineligible for matching, effectuating simplified balanced risk set matching.[22] To assess matching performance, propensity score distributions for each group were examined once more, and LOS was analyzed similarly to the original five case definitions.

[22] Li, Propert, and Rosenbaum (2001), "Balanced Risk Set Matching".

To further characterize dependence of LOS on the time of CDI onset, we fit a nonparametric multistate model consistent with previous studies.[23] The model has two transient states (admitted-uninfected and admitted-infected) and one absorbing state (discharged); each patient starts in the admitted-uninfected state. Figure 5.12 is a diagram of all allowed transitions. For all case definitions with a diagnosis time [(ii), (iii), and (iv)], an Aalen-Johansen estimator was used on the full, unmatched dataset to calculate time-varying hazards of each transition. Since the estimator is sensitive to regions with sparsity or outliers, the minimum LOS was specified as 1 day, the time of CDI diagnosis was left- or right-shifted to at least 0.5 days from admission and discharge events, and

[23] Mitchell et al. (2014); Stevens et al. (2015); Kleef et al. (2014), "Excess length of stay and mortality due to Clostridium difficile infection: A multi-state modelling approach".

the model was computed to a precision of 0.1 day up to the 99th percentile LOS value [38.9, 41.9, and 40.9 days for case definitions (ii), (iii), and (iv), respectively]. The mean excess LOS was then estimated as the average difference in LOS between patients with and without CDI at each time t, weighted by the distribution of times spent in the uninfected state. Robust 95% CIs were generated from 1,000 bootstrap resampling runs.

Analyses were performed in R 3.2.2 (R Foundation for Statistical Computing, Vienna, Austria) using the `glmnet`,[24] `ROCR`,[25] `MatchIt`,[26] `survival`, and `etm`[27] packages. All software code for the analysis is available at: `https://github.com/powerpak/cdi-cost`

[24] Friedman, Hastie, and Tibshirani (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent".

[25] Sing et al. (2005), "ROCR: Visualizing classifier performance in R".

[26] Ho et al. (2011), "MatchIt : Nonparametric Preprocessing for Parametric Causal Inference".

[27] Allignol, Schumacher, and Beyersmann (2011), "Empirical Transition Matrix of Multi-State Models: The etm Package".

Results

371,622 records of visits during the study time range were queried from the EMR, with 23,968 variables extracted for each visit (Figure 5.3). After filtering for the index visit per adult patient and excluding deaths and invalid dates, 171,938 visits were eligible for inclusion and classified into five overlapping case definitions for CDI. Case cohort sizes before matching and overlaps are depicted in Figure 5.4.

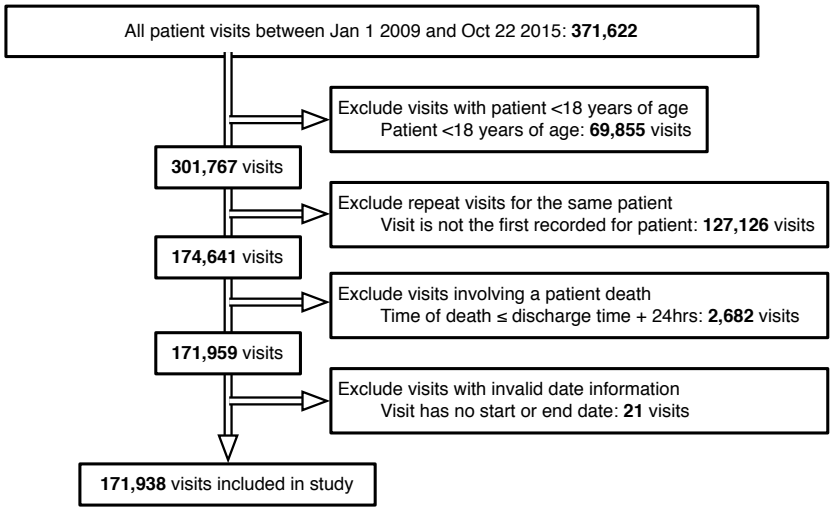


Figure 5.3: **Inclusion/exclusion procedure for the present study.** Double-line arrows indicate the procession of visit records.

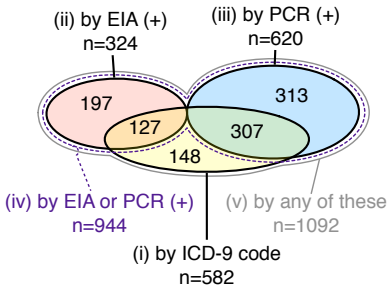


Figure 5.4: **Cohort sizes for each case definition and cohort intersections before matching.** Venn diagram of case cohort sizes for each of the five *C. difficile* infection case definitions, before matching, with sizes of all intersections (overlaps) between case definitions. Areas are not to scale. There is no intersection between case definitions (ii) and (iii), since only the first positive toxin assay result for each visit was examined. Case definition (iv), "by EIA or PCR (+)," is a strict superset of case definitions (ii) and (iii). Case definition (v), "by any of these," is a strict superset of case definitions (i), (ii), and (iii). Sizes of matched case cohorts are provided in Table 5.1.

Regularized logistic regression models predicting the risk of CDI acquisition were fit to EMR data from the first 24 hours of each admission for each case definition. The cross-validated mean area under the receiving operator charac-

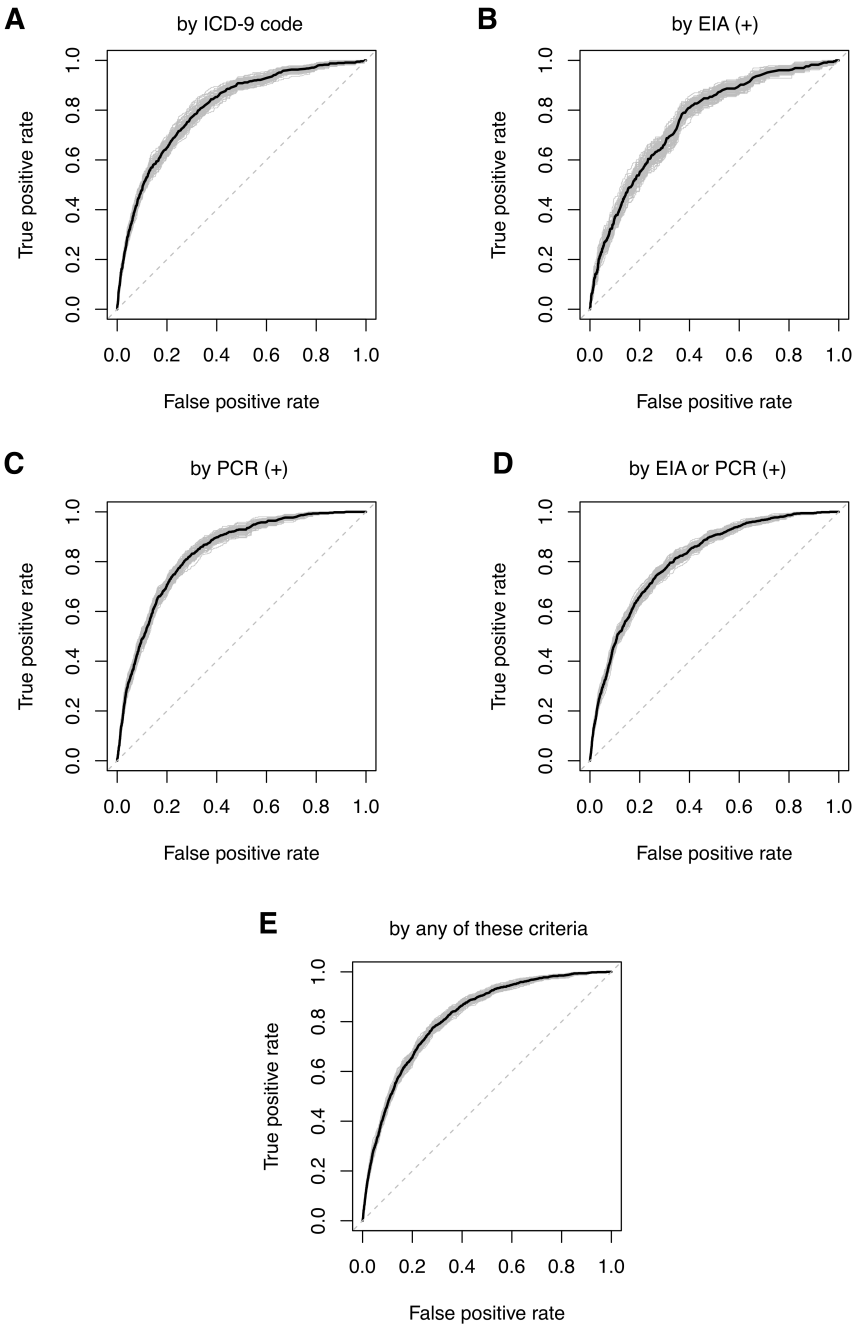


Figure 5.5: **Receiver operator characteristic curves for *C. difficile* infection propensity models.** A-E, receiver operator characteristic (ROC) curve comparing the true positive rate against the false positive rate for every possible cutoff of the propensity score, for each of the five *C. difficile* infection case definitions. ROCs are only measured on test data not used to train the models, under five-fold cross validation. Light grey lines indicate ROCs for 100 bootstrap samples. Area under each ROC is reported in the main text under Results. ICD-9, International Classification of Diseases Ninth Revision; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

teristic (AUROC), measuring performance of each model, varied only slightly between case definitions (Figure 5.5): (i) by ICD-9 code, 0.81 (95% confidence interval [CI], 0.79–0.83); (ii) by positive toxin EIA, 0.76 (95% CI, 0.74–0.78), (iii) by positive toxin PCR, 0.84 (95% CI, 0.82–0.85), (iv) by either toxin assay, 0.81 (95% CI, 0.80–0.82); (v) by any of these, 0.82 (95% CI, 0.81–0.83). The number of selected variables in each model ranged from 373 to 1,027.

		Matched cohorts for each CDI case definition					
		(i) by ICD-9 code			(ii) by EIA (+)		
	No. (%)	No. (%)			No. (%)		
Characteristic	All patients (n=171,936)	All controls (n=171,356)	Matched cases & controls[a] (n=489)	SMD after matching (*P* value)	All controls (n=73,647)	Matched cases & controls[a] (n=274)	SMD after matching (*P* value)
Female sex	101,964 (59)	101,638 (59)	278 (57)	0 (1)	44,132 (60)	145 (53)	0 (1)
Age[b]							
18-29	22,344 (13)	22,266 (13)	69 (14)		9,552 (13)	22 (8)	
30-44	39,003 (23)	38,898 (23)	86 (18)		16,451 (22)	26 (9)	
45-59	37,234 (22)	37,129 (22)	90 (18)	0.016 (0.86)	15,956 (22)	58 (21)	0.018 (0.79)
60-74	43,946 (26)	43,802 (26)	122 (25)		18,407 (25)	83 (30)	
75-90	26,167 (15)	26,041 (15)	106 (22)		11,817 (16)	70 (26)	
≥90	3,244 (2)	3,220 (2)	16 (3)		1,464 (2)	15 (5)	

	Matched cohorts for each CDI case definition (cont.)								
	(iii) by PCR (+)			(iv) by EIA or PCR (+)			(v) by any of the criteria		
	No. (%)			No. (%)			No. (%)		
Characteristic	All controls (n=97,351)	Matched cases & controls[a] (n=493)	SMD after matching (*P* value)	All controls (n=170,994)	Matched cases & controls[a] (n=788)	SMD after matching (*P* value)	All controls (n=170,846)	Matched cases & controls[a] (n=945)	SMD after matching (*P* value)
Female sex	57,340 (59)	254 (52)	0 (1)	101,469 (59)	408 (52)	0 (1)	101,390 (59)	493 (52)	0 (1)
Age[b]									
18-29	12,714 (13)	47 (10)		22,265 (13)	79 (9)		22,245 (13)	87 (9)	
30-44	22,430 (23)	72 (15)		38,879 (23)	124 (13)		38,845 (23)	134 (14)	
45-59	21,069 (22)	117 (24)	0.005	37,025 (22)	209 (23)	0.003	36,999 (22)	208 (22)	0.004
60-74	25,273 (26)	136 (28)	(0.99)	43,680 (26)	266 (29)	(0.98)	43,643 (26)	267 (28)	(0.93)
75-90	14,120 (15)	114 (23)		25,936 (15)	231 (24)		25,912 (15)	217 (23)	
≥90	1,745 (2)	7 (1)		3,209 (2)	35 (3)		3,202 (2)	32 (3)	

Table 5.1: **Demographic characteristics of the study population and matched cohorts.** Abbreviation: CDI, Clostridium difficile infection; ICD-9, International Classification of Diseases Ninth Revision; EIA, enzyme immunoassay; PCR, polymerase chain reaction; SMD, standardized mean difference.

[a] Separate columns are unnecessary because 1:1 exact matching was performed on the characteristics shown, and therefore all values are identical.

[b] SMD is shown for age treated as a continuous variable; coarsened exact matching was performed using the listed age ranges.

For each case definition, over 75% of cases were successfully matched by propensity score to controls (Figure 5.4 and Table 5.1). The groups are well matched on demographics and propensity scores (all *P* values >0.1 and standardized differences between –0.1 and 0.1; propensity score distributions in Figure 5.6). Differences in the median LOS between matched case and control cohorts for all CDI case definitions were strongly statistically significant, although

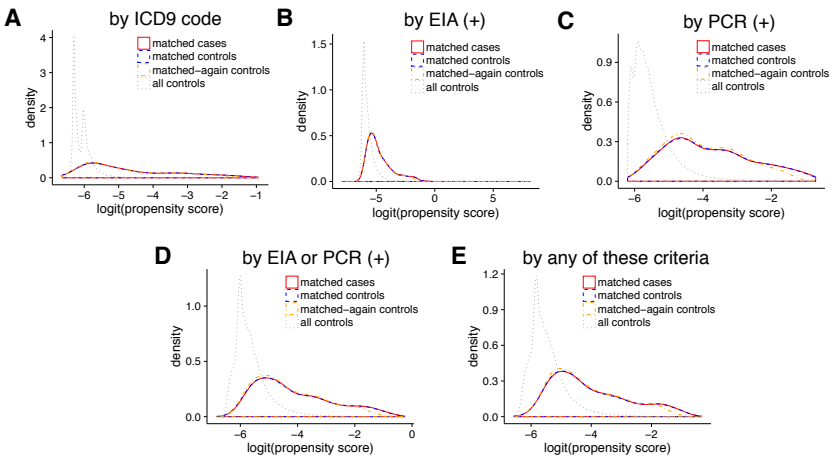


Figure 5.6: **Propensity score distributions for matched cohorts for each *C. difficile* infection case definition.** A-E, density plots of propensity score distributions for matched cases, matched controls, matched-again controls, and all controls for each of the five CDI case definitions. Matched-again controls are derived from a second round of matching between the case-matched controls from the first round of matching and remaining unmatched controls. All X axes are logit-scaled. All Y axes are scaled to unit probabilities; the area under every curve equals 1. The matching algorithm intends to align the propensity score distributions for all of the matched groups.

the magnitude of the differences varied greatly between definitions (Figure 5.7). The differences in the median LOS by case definition were: (i) by ICD-9 code, 3.1 days (95% CI, 2.2–3.9); (ii) by positive toxin EIA, 10.1 days (95% CI, 7.3–12.2), (iii) by positive toxin PCR, 6.6 days (95% CI, 5.0–8.1), (iv) by either toxin assay, 7.2 days (95% CI, 5.8–8.3); and (v) by any of these, 5.7 days (95% CI, 4.5–6.6). There were no significant differences in LOS for a second round of

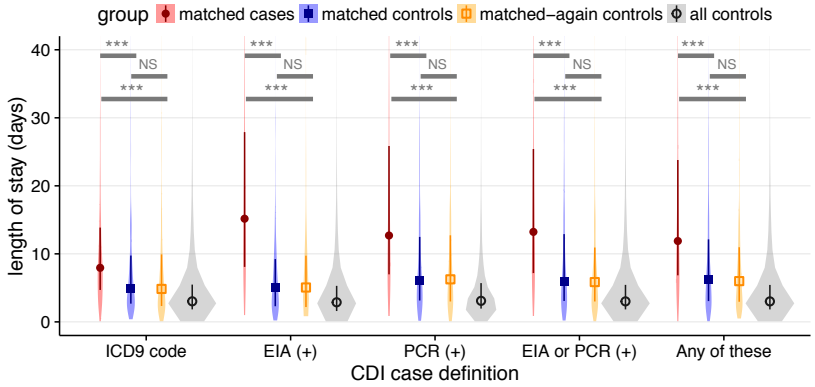


Figure 5.7: **Changes in length of stay for five case definitions of *C. difficile* infection, not accounting for time of infection.** Violin plots of the distributions in length of stay for matched cases, matched controls, matched-again controls, and all controls, for each of the five case definitions. Darker points and vertical bars depict the median and interquartile range for each group. Horizontal bars depict Mann-Whitney *U* tests for significance of differences between groups (***, Bonferroni-corrected $P < 0.001$; NS, not significant [$P > 0.1$]). CDI, *C. difficile* infection.

matching between matched controls and remaining controls (matched-again controls) for any of the case definitions (Figure 5.7). Kaplan-Meier curves for the time-dependent risk of being discharged from the hospital showed significant differences between matched case and control cohorts up to post-admit day 60 for all case definitions except ICD-9 code (Figure 5.8).

Estimates of LOS associated with CDI are inflated by dependencies on time-

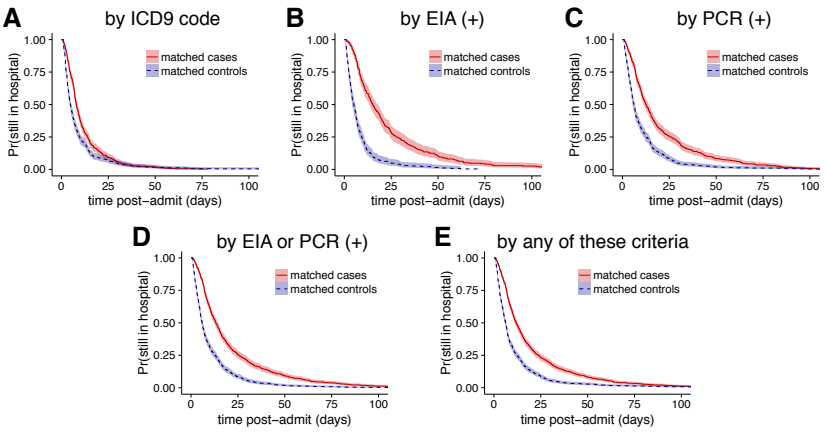


Figure 5.8: **Kaplan-Meier plots of the time-dependent probability for a patient to still be in the hospital, not accounting for time of infection.** A-E, Kaplan-Meier plots of the time-dependent probability for a patient to still be in the hospital, comparing matched cases and controls for each case definition of *C. difficile* infection. Shaded areas depict 95% confidence intervals calculated from standard errors. ICD9, International Classification of Diseases Ninth Revision; EIA, enzyme immunoassay; PCR, polymerase chain reaction.

to-infection—if longer pre-infection LOS increases CDI risk, i.e., reverse causation, this leads to overestimates in attributable cost.[28] We therefore performed two follow-up analyses to account for this. First, we stratified the LOS comparison by the time of CDI diagnosis for case definition (iv) into 0–3 day, 3–8 day, and ≥8 day case cohorts, training new propensity models for another matched comparison, with similar matching performance (Figure 5.9). Since 3 days (or

[28] Mitchell et al. (2014); Stevens et al. (2015).

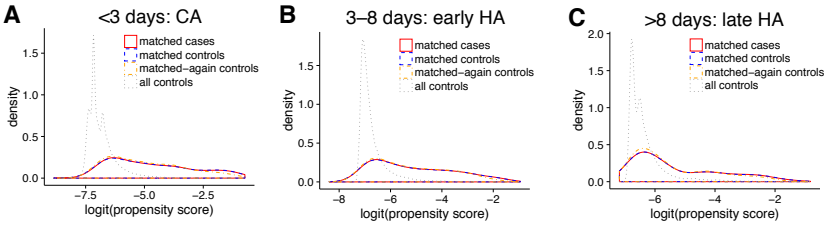


Figure 5.9: **Propensity score distributions for matched cohorts stratified by time of CDI diagnosis.** A-C, density plots of propensity score distributions for matched cases, matched controls, matched-again controls, and all controls for cases defined by any positive toxin assay, stratified by the time to infection. Matched-again controls are derived from a second round of matching between the case-matched controls from the first round of matching and remaining unmatched controls. All X axes are logit-scaled. All Y axes are scaled to unit probabilities; the area under every curve equals 1. The matching algorithm intends to align the propensity score distributions for all of the matched groups. CA, community acquired; HA, health- care associated.

72 hours) is a typical cutoff for differentiating community acquired (CA) from healthcare-associated (HA) CDI,[29] these strata were named "CA," "early HA," and "late HA," respectively. As suspected, stratification revealed a positive correlation between time of diagnosis and the CDI-associated difference in LOS (Figure 5.10). The differences in medians were: for CA, 2.5 days (95% CI, 1.2–3.4); early HA, 3.1 days (95% CI, 1.8–4.4); and late HA, 14.0 days (95% CI, 9.9–17.1). All comparisons between matched cases and controls were again strongly statistically significant, and comparisons between matched controls and matched-again controls were not significant (Figure 5.10). Kaplan-Meier plots likewise confirmed a correlation between time of CDI diagnosis and the difference in

[29] Longtin et al. (2016), "Effect of Detecting and Isolating Clostridium difficile Carriers at Hospital Admission on the Incidence of C difficile Infections: A Quasi-Experimental Controlled Study."; Polage et al. (2015).

time-dependent discharge risk (Figure 5.11).

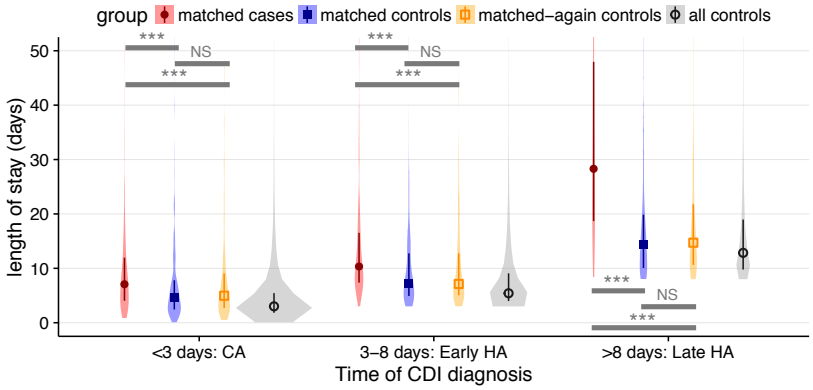


Figure 5.10: **Changes in length of stay for *C. difficile* infection defined by any positive toxin assay, stratified by the time to infection.** Violin plots of the distributions in length of stay for matched cases, matched controls, matched-again controls, and all controls, for three ranges of the result time for the first positive toxin assay. Points and vertical bars depict the median and interquartile range for each group. Horizontal bars depict Mann-Whitney *U* tests for significance of differences between groups (***, Bonferroni-corrected $P < 0.001$; NS, not significant [$P > 0.1$]). CA, community acquired; HA, healthcare associated.

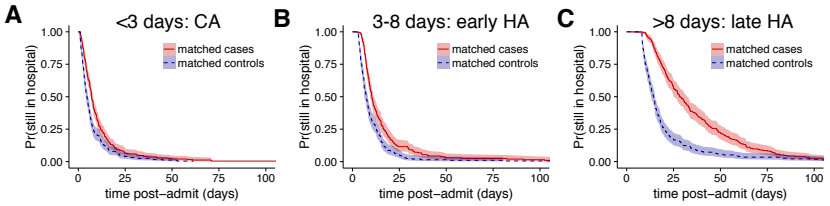


Figure 5.11: **Kaplan-Meier plots of the time-dependent probability for a patient to still be in the hospital, stratifying patients by the time to infection.** A-C, comparison of matched *C. difficile* infection cases and controls for the same three ranges of the time of the first positive toxin assay, stratified by the time to infection. Shaded areas depict 95% confidence intervals calculated from standard errors. CA, community acquired; HA, healthcare associated.

To further address reverse causation, we fit a multistate model similar to previously published studies[30] that explicitly estimates time-dependent, competing risks of transitioning to a CDI-positive state vs. being discharged. Figure 5.12 depicts the three states and allowed transitions of the model. After fitting the model for the case definitions with a time of diagnosis (ii, iii, and iv), the expected remaining LOS can be compared across cohorts that have already transitioned to the CDI infected state vs. those that are still CDI negative at any given timepoint (Figure 5.13).

To summarize the overall relationship between CDI and LOS, differences in LOS were weighted by the distribution of times spent in the initial state and averaged. The average differences for each case definition were: (ii) by positive toxin EIA, 3.0 days (95% CI, 2.0–4.0); (iii) by positive toxin PCR, 3.5 days (95% CI, 2.7–4.5); and (iv) by either toxin assay, 3.3 days (95% CI, 2.6–4.0). Notably, the 95% CI for the difference in cohort (iv) overlaps the 3.1 day difference for the "early HA" stratum of the propensity-matched analysis in the same cohort.

[30] Mitchell et al. (2014); Stevens et al. (2015); Kleef et al. (2014).

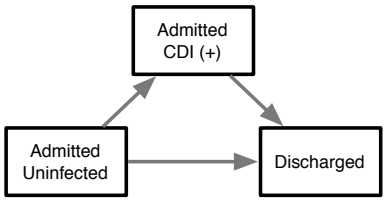


Figure 5.12: **Multistate model of *C. difficile* infection.** Three states of the multistate model and allowed transitions. Patients may only transition in the direction of the arrows.

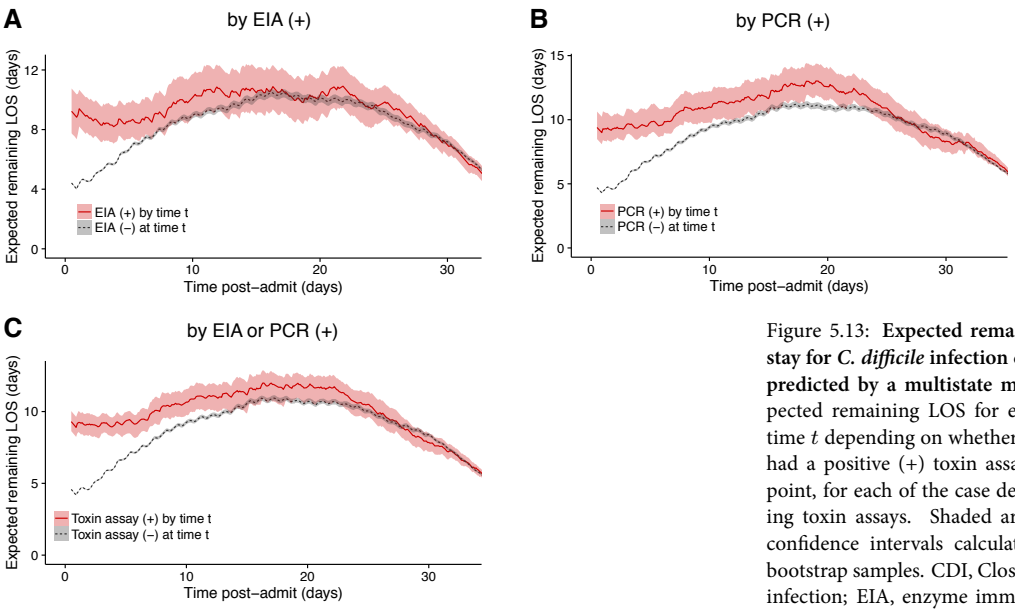


Figure 5.13: **Expected remaining length of stay for *C. difficile* infection case definitions predicted by a multistate model.** A-C, expected remaining LOS for each post-admit time t depending on whether the patient has had a positive (+) toxin assay by that timepoint, for each of the case definitions involving toxin assays. Shaded areas depict 95% confidence intervals calculated from 1,000 bootstrap samples. CDI, Clostridium difficile infection; EIA, enzyme immunoassay; PCR, polymerase chain reaction; LOS, length of stay.

Discussion

This study examined nearly seven years of uncurated EMR data for a single hospital and determined associated costs of CDI as defined by either visit diagnosis codes or lab results. In the analysis unadjusted for time-to-infection, differences in LOS were often greater than national averages from similar unadjusted studies,[31] but changes in the case definition resulted in substantial changes in the estimated differences in LOS. Although two hospitals reported good concordance between ICD-9 codes and CDI toxin assay results,[32] this is not necessarily the case for all hospitals. We found that 75% of ICD-9 coded visits involved a positive toxin assay, while only 46% of visits with a positive toxin assay had the ICD-9 code (Figure 5.4). Changes in LOS were not significantly different between EIA and PCR toxin assays, although our study was limited by a smaller sample size for EIA (+) cases. Toxin assays are likely a more reliable CDI definition given their basis in clinical symptoms and evidence for CDI, whereas medical coding suffers from biases introduced by billing and reimbursement.[33]

Treating CDI as a baseline condition by ignoring the relationship between pre-infection hospital exposure and CDI risk overestimates associated costs.[34] Unlike visit diagnosis codes, toxin assay results provide a presumptive time-to-

[31] Gabriel and Beriot-Mathiot (2014); Zhang et al. (2016); Zimlichman et al. (2013).

[32] Dubberke et al. (2006), "ICD-9 codes and surveillance for Clostridium difficile-associated disease"; Scheurer et al. (2007), "Accuracy of ICD-9 coding for Clostridium difficile infections: a retrospective cohort."

[33] Rhee et al. (2015), "Improving documentation and coding for acute organ dysfunction biases estimates of changing sepsis severity and burden: a retrospective study"; Romano and Mark (1994), "Bias in the coding of hospital discharge data and its implications for quality assessment."

[34] Graves et al. (2010), "Estimating the cost of health care-associated infections: mind your p's and q's."; Mitchell et al. (2014); Stevens et al. (2015).

infection that we incorporated into two different statistical methods addressing time-dependent bias. When using a case definition of either toxin assay being positive, the measured difference in LOS in the multistate model corresponded closely with the difference seen in the "early HA" stratum of a time-stratified propensity-matched analysis (3.3 vs. 3.1 days). This suggests that measured differences in this study robustly reflect associated costs of HA-CDI in our patient population. Since estimates for each time-to-infection stratum in the matching analysis differed greatly (Figure 5.10–5.11), time-to-infection clearly contributed bias to the unstratified analysis (Figure 5.7–5.8), demonstrating how the many studies that ignore this bias[35] produce inflated estimates. In our dataset, ignoring time-dependent bias would lead to a more than two-fold overestimation of CDI-associated LOS. Given our findings, we cautiously interpret the results of meta-analyses that conflate ICD-9 code and toxin assay case definitions and often ignore time-dependent bias.[36]

[35] Gabriel and Beriot-Mathiot (2014); Zhang et al. (2016); Zimlichman et al. (2013).

[36] Gabriel and Beriot-Mathiot (2014); Ghantoji et al. (2010); Zhang et al. (2016).

To our knowledge, this is the first study that uses machine learning on uncurated EMR data to estimate the local cost of CDI. Our models of CDI risk performed on par with prior models fitted to lower-dimensional data.[37] Since our models are based on tens of thousands of structured fields in the EMR that require neither chart review nor manual curation beyond masking known CDI-related effects, re-analysis of future data is inexpensive. Starting from exported visit data, the entire analysis runs in several hours on standard desktop computers. Therefore, the effects of new interventions against CDI can be efficiently monitored over time, e.g., continually testing whether new treatments actually lower the CDI-associated LOS or quantifying cost savings of new preventive strategies that decrease CDI incidence. Changes in LOS can be extrapolated to approximate economic costs by multiplying by the average cost of extra inpatient-days as LOS is the main contributor to CDI's cost.[38] In our dataset, using the time-dependency adjusted differences in LOS of 3.1–3.3 days and the national average cost of additional inpatient-days for CDI cases,[39] the median cost associated with each case would be approximately $10,600–11,300; this is substantial in comparison to the national average price for an inpatient visit—approximately $13,000 in 2011.[40] Using the average yearly caseload observed in the dataset for toxin assay positive cases, our figures represent an annual accounting cost to Mount Sinai of approximately $1.5 million, not including the opportunity cost of bed occupancy by CDI patients or the impact on infection

[37] Dubberke et al. (2011), "Development and validation of a Clostridium difficile infection risk prediction model."; Tanner et al. (2009), "Waterlow score to predict patients at risk of developing Clostridium difficile-associated disease"; Wiens, Guttag, and Horvitz (2014).

[38] Graves et al. (2010); McGlone et al. (2012); Wilcox et al. (1996); Zimlichman et al. (2013).

[39] Zimlichman et al. (2013).

[40] Cooper et al. (2015).

control resources.[41] In principle, our analysis is generalizable to any HAI where lab results recorded in the EMR robustly reflect the incidence of infections.

Our study has several limitations. The analysis was designed with conservative assumptions, preferring that the models underestimate rather than overestimate changes associated with CDI. For example, restricting to one index visit per patient certainly excluded many repeat visits for recurrent CDI, which are known to incur higher costs.[42] We preferred a relatively simple machine learning technique, elastic net regularized generalized linear models, because of its transparency in variable selection and unparalleled speed on sparse training data. More advanced techniques might marginally improve propensity model accuracy at the expense of computational complexity and model interpretability. EMR data has known drawbacks compared to clinical research data, such as limitations in time precision, the sparsity of the data, and increased opportunity for coding error. Also, retrospective matching-based studies face known trade-offs in estimating actual effects of HAIs,[43] so we used two separate statistical approaches to control for time-dependent bias, ultimately finding congruous results. We did not have structured billing data, so we cannot trace itemized costs and characterize the exact relationship between LOS and costs beyond the proportional estimate above. Finally, only one hospital's data was used to implement this analysis, which would benefit from comparison to other hospitals' data. Therefore, we provide complete code for our analysis so that it may be re-implemented elsewhere and improved by the community.

[41] Graves et al. (2010).

[42] Dubberke et al. (2008); Dubberke et al. (2014); Rodrigues, Barber, and Ananthakrishnan (2016), "A Comprehensive Study of Costs Associated With Recurrent Clostridium difficile Infection".

[43] Graves et al. (2010).

Conclusions

Two independent statistical analyses adjusting for time-dependent bias produced similar results for the CDI-associated change in LOS at Mount Sinai (3.1 and 3.3 days), suggesting that automated methods based on machine learning and uncurated EMR data robustly and conservatively estimate the local cost of an HAI in both LOS and financial terms. This procedure is transparent, reproducible, and inexpensive, suggesting that hospitalists and infection control officers can leverage EMR data to estimate their specific, local costs of HAIs on an ongoing basis rather than relying on widely varying benchmarks published by other institutions.

Notes

Contributors

Theodore Pak (TRP), Kieran Chacko (KC), Timothy O'Donnell (TO), Shirish Huprikar (SH), Harm van Bakel (HVB), Andrew Kasarskis (AK), and Erick R. Scott (ERS) contributed to this study. TRP, KC, AK, and ERS designed the study and developed the models. TRP, KC, TO, SH, HVB, AK, and ERS acquired the data. TRP performed the data analysis and created all figures in this chapter. SH reviewed selected variables for the propensity model. TO provided technical support, and HVB, SH, and AK provided administrative and material support. AK and ERS supervised the study. TRP wrote the first draft of this chapter. TRP is the first author on the corresponding submitted manuscript.

Funding

This work was supported by the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai, and also in part by the NIH/NIAID grants F30-AI122673 (TRP), T32-GM007280 (TRP), and R01-AI119145 (HVB, KC). The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

This work was supported in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Deena Altman, Camille Hamula, and Gopi Patel for their assistance in improving the design of the study and reviewing the manuscript.

6

Mass cytometry and transcriptomics of the innate immune response to chikungunya infection reveals a central role for monocytes

Chikungunya virus (CHIKV) is a globally epidemic mosquito-borne alphavirus causing acute and chronic arthritic disease. Our study used a systems immunology approach by incorporating whole blood RNA-seq, 35-plex mass cytometry of PBMCs, and serum cytokine measurements of acute and convalescent phase samples from 42 natural pediatric infections in Nicaragua. Semi-supervised classification and clustering of single-cell events into 57 sub-communities of canonical leukocyte phenotypes revealed a monocyte-driven response to acute infection, with greatest expansions of "intermediate" $CD14^{++}CD16^{+}$ monocytes and an activated subpopulation of $CD14^{+}$ monocytes. Increases in acute phase CHIKV surface protein expression were highest for monocytes and dendritic cells, although surprisingly, B cell subpopulations also displayed significant increases. Serum cytokine measurements confirmed significant acute phase upregulation of monocyte chemoattractants. Transcriptomic signatures were revealed not only for infection phase, but also for convalescent phase immunogenicity, acute phase viremic load, and symptom severity. Finally, we present a multiscale network that summarizes all observed modulations across cellular and gene expression levels and their interactions with clinical outcomes, providing a uniquely global view of the biomolecular landscape of CHIK pathophysiology.

I'd like to share a revelation that I've had during my time here. It came to me when I tried to classify your species and I realized that you're not actually mammals. Every mammal on this planet instinctively develops a natural equilibrium with the surrounding environment but you humans do not. You move to an area and you multiply and multiply until every natural resource is consumed and the only way you can survive is to spread to another area. There is another organism on this planet that follows the same pattern. Do you know what it is? A virus. Human beings are a disease, a cancer of this planet. You're a plague and we are the cure.

—AGENT SMITH, *The Matrix*

I pictured myself as a virus, or a cancer cell, and tried to sense what it would be like.

—JONAS SALK

CHIKUNGUNYA (CHIKV) is a re-emerging mosquito-borne alphavirus that causes endemic and explosively epidemic infections throughout tropical regions of the world.[1] Transmittable by *Aedes egypti* and *Aedes albopictus*, the same vectors as Dengue virus (DENV), phylogenies of CHIKV have indicated

[1] Weaver and Lecuit (2015), "Chikungunya virus and the global spread of a mosquito-borne disease."

that urban endemic strains originated from several transmission events out of enzootic, sylvatic cycles between non-human primates and arboreal mosquitos in eastern Africa.[2] In 2004, the largest outbreak ever recorded spread rapidly from Africa through the Indian Ocean and Asia to Papua New Guinea and islands in the Pacific; subsequently, the first autochthonus transmissions in the Americas and the US were reported in 2013.[3] Millions of cases have now been reported in at least forty countries.[4]

Unlike other arboviral diseases like DENV, the majority of CHIKV infections produce symptomatic illness, with typical manifestations consisting of fever, a diffuse body rash, and joint pain and inflammation.[5] Notably, debilitating joint-related symptoms can persist for years mimicking rheumatoid arthritis in up to 50% of afflicted populations—the namesake characteristic of the disease, as "chikungunya" is a Swahili/Makonde word describing a bent posture.[6] Rarely, complications can occur, including encephalopathy and encephalitis, fulminant hepatitis, and myocarditis.[7] Mortality occurs in approximately 0.1% of cases.[8] Besides anti-inflammatories for symptomatic relief, there are no specific treatments available for CHIKV.[9] Several vaccine candidates have reached preclinical or phase I trials,[10] but major commercial investment will be required to complete their development, and finding clinical sites to demonstrate efficacy will be complex because of the unpredictable incidence and spread of the virus.[11] Until a vaccine or antiviral agent is available, prevention efforts will remain focused on mosquito control.[12]

Because of CHIKV's recent re-emergence in the Western hemisphere, there are profound gaps in the understanding of CHIKV pathogenesis, including uncertainty over the roles of viral proteins and the myriad genetic and signaling factors involved in a successful or unsuccessful immune response,[13] particularly in pediatric cases.[14] CHIKV can infect many cell types, including skin fibroblasts, endothelial cells, primary epithelia, and human muscle satellite cells.[15] Reports of tropism in subsets of peripheral blood monocytes (PBMCs) vary, with CHIKV antigens detected in monocytes during acute infection,[16] although primary monocytes and macrophages only appear to be infectable at low efficiency in vitro.[17] Thus far, primary B and T cells have not been successfully infected in vitro.[18] Monocytes and macrophages are thought to have a substantial role in the acute inflammatory response to CHIKV, as primate models show consistent recruitment of these cell types and natural killer (NK) cells

[2] Volk et al. (2010), "Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent Epidemics and Various Evolutionary Rates".

[3] Nasci (2014), "Movement of chikungunya virus into the Western hemisphere."

[4] Suhrbier, Jaffar-Bandjee, and Gasque (2012), "Arthritogenic alphaviruses: an overview".

[5] Couderc and Lecuit (2015), "Chikungunya virus pathogenesis: From bedside to bench"; Weaver and Lecuit (2015).

[6] Miner et al. (2015), "Brief report: Chikungunya viral arthritis in the United States: A mimic of seronegative rheumatoid arthritis"; Weaver and Lecuit (2015).

[7] Rolph, Foo, and Mahalingam (2015), "Emergent chikungunya virus and arthritis in the Americas".

[8] Ibid.

[9] Suhrbier, Jaffar-Bandjee, and Gasque (2012); Weaver and Lecuit (2015).

[10] Plante et al. (2015), "Extended Preclinical Safety, Efficacy and Stability Testing of a Live-attenuated Chikungunya Vaccine Candidate"; Weger-Lucarelli et al. (2014), "A Novel MVA Vectored Chikungunya Virus Vaccine Elicits Protective Immunity in Mice."

[11] Weaver and Lecuit (2015).

[12] Ibid.

[13] Assunção-Miranda, Cruz-Oliveira, and Da Poian (2013), "Molecular mechanisms involved in the pathogenesis of alphavirus-induced arthritis."; Sourisseau et al. (2007), "Characterization of reemerging chikungunya virus."; Weaver and Lecuit (2015)

[14] Teng et al. (2015), "A Systematic Meta-analysis of Immune Signatures in Patients With Acute Chikungunya Virus Infection".

[15] Couderc and Lecuit (2015); Lum and Ng (2015), "Cellular and molecular mechanisms of chikungunya pathogenesis".

[16] Her et al. (2010), "Active infection of human blood monocytes by Chikungunya virus triggers an innate immune response".

[17] Sourisseau et al. (2007); Teng et al. (2012), "Viperin restricts chikungunya virus replication and pathology".

[18] Her et al. (2010); Sourisseau et al. (2007); Teng et al. (2012).

into infected tissues,[19] and mouse models treated with bindarit (an inhibitor of monocyte chemoattractant CCL2) showed reduced monocyte recruitment, joint swelling, and bone loss following infection.[20] The role of monocytes appears to be protective as well as inflammatory. For example, mice deficient for CCR2 (the receptor for CCL2) show prolongation of arthritic disease corresponding with replacement of the monocyte/macrophage infiltrate in infected joints by neutrophils and eosinophils.[21] In humans, however, details of the relationship between monocyte subpopulations, acute phase pathogenesis, and chronic symptomatology remain poorly understood.[22]

The innate immune response, particularly via type I interferon (IFN) signaling, is important for control of CHIKV replication during the acute phase of infection.[23] CHIKV infection acutely induces high levels of IFNα release in both humans and model organisms.[24] In mouse models, type I IFNs control CHIKV replication by directly acting on nonhematopoietic cells, likely via activation of host sensors for viral RNA, such as RIG-I and MDA5.[25] Additionally, either IRF-3 or IRF-7 signaling appears to be independently sufficient for preventing lethality of CHIKV infection in adult mice.[26] In primary cell culture and mice, interferon stimulated genes such as the OAS family and RSAD2 (Viperin) appear to exert antiviral roles against CHIKV, although the details of these signal transduction pathways and their relative importance are unresolved.[27]

[19] Labadie et al. (2010), "Chikungunya disease in nonhuman primates involves long-term viral persistence in macrophages."

[20] Chen and Shapiro (2015), "The advent of genome-wide association studies for bacteria"; Rulli et al. (2011), "Protection from arthritis and myositis in a mouse model of acute chikungunya virus disease by bindarit, an inhibitor of monocyte chemotactic protein-1 synthesis".

[21] Poo et al. (2014), "Multiple Immune Factors Are Involved in Controlling Acute and Chronic Chikungunya Virus Infection".

[22] Burt et al. (2017), "Chikungunya virus : an update on the biology and pathogenesis of this emerging pathogen"; Weaver and Lecuit (2015).

[23] Burt et al. (2017); Schilte et al. (2010), "Type I IFN controls chikungunya virus via its action on nonhematopoietic cells."

[24] Labadie et al. (2010); Teng et al. (2015).

[25] Schilte et al. (2010).

[26] Schilte et al. (2012), "Cutting edge: independent roles for IRF-3 and IRF-7 in hematopoietic and nonhematopoietic cells during host response to Chikungunya infection."

[27] Burt et al. (2017).

Historically, the immune system has been described by evaluating individual components in isolation. This approach is often biased toward better-recognized phenotypes and pathways and is likely to miss globally significant patterns of interconnectivity, particularly across the multiple conjoint scales of the immune system, e.g., transcriptional modulation within cells, resultant expansion and contraction of certain cell populations, and crosstalk between those immune cells and disparate tissues.[28] Genome-wide expression profiling using microarrays or RNA-seq and mass cytometry using Cytometry by Time-of-Flight (CyTOF) offer the capability to perform unbiased, systematic exploration of hundreds of thousands of changes transpiring within a particular perturbation of the immune system. Weighted coexpression and probabilistic causal network models can then synthesize data from "omic" assays into a map of quantitative relationships between all regulatory elements of a particular immune response, which is one of the goals of systems immunology.[29]

[28] Kidd et al. (2014), "Unifying immunology with informatics and multiscale biology."

[29] Arazi et al. (2013), "Human systems immunology: Hypothesis-based modeling and unbiased data-driven approaches"; Germain et al. (2011), "Systems biology in immunology: a computational modeling perspective."

Although biomolecular network models have demonstrated utility in finding causal gene modules and novel mechanisms for complex, inheritable human diseases,[30] because of the difficulty in acquiring data at the scale necessary for fitting these models, they remain relatively new in the field of infectious diseases. Even still, network models have already helped map detailed regulatory circuits in hematopoiesis, transcriptional regulation of hundreds of leukocyte populations in mice, and viral sensing mechanisms in dendritic cells via Toll-like receptors (TLRs).[31]

Previous observational studies of the immune response to CHIKV in natural human infections typically concentrated on protein or gene expression levels of a small number of cytokines and inflammatory mediators,[32] often producing conflicting results.[33] Among studies that used cytometry, one group employed CyTOF to profile ten CHIKV patients, but their analysis focused almost entirely on T cells.[34] Our study, by contrast, employs a systems immunology approach, integrating three high-throughput techniques to comprehensively profile the acute and convalescent phases of 42 pediatric cases of natural CHIKV infection in Managua, Nicaragua. To our knowledge, we provide here the first published RNA-seq study of CHIKV infection in humans, the first report of CHIKV-induced modulations for nearly all peripheral blood mononuclear cell (PBMC) subpopulations in humans, and the only study that applies multiple molecular profiling techniques to pediatric cases of CHIKV. We used whole blood samples from patients visiting an acute care facility that were lab-confirmed as cases of CHIKV viremia, comparing each patient's acute phase sample (1-2 days post symptom onset [p.s.o.]) against paired samples taken two weeks later, after resolution of symptoms and viremia. We analyzed these samples using CyTOF, RNA-seq, and multiplex ELISA (Luminex), employing a systematic, hypothesis-free approach for finding globally significant changes in cell subpopulation frequencies, gene expression, and serum cytokine concentrations during acute CHIKV infection. We then searched for interactions within all of these data and with measurements of clinical outcomes such as the viral titer during the acute phase, the severity of symptoms at presentation, and the convalescent phase CHIKV immunoglobulin G (IgG) titer. Finally, to synthesize these interactions across the three examined scales, we present a multiscale network model that summarizes all correlations between gene modules, cell subpopulations, and clinical variables in this study.

[30] Chen et al. (2008), "Genetic Compatibility and Virulence of Reassortants Derived from Contemporary Avian H5N1 and Human H3N2 Influenza A Viruses"; Emilsson et al. (2008), "Genetics of gene expression and its effect on disease."; Huan et al. (2015), "Integrative network analysis reveals molecular mechanisms of blood pressure regulation"; Zhang et al. (2013), "Glycosylation on Hemagglutinin Affects the Virulence and Pathogenicity of Pandemic H1N1/2009 Influenza A Virus in Mice"

[31] Kidd et al. (2014).

[32] Chaaitanya et al. (2011), "Role of proinflammatory cytokines and chemokines in chronic arthropathy in CHIKV infection."; Chow et al. (2011), "Persistent arthralgia induced by Chikungunya virus infection is associated with interleukin-6 and granulocyte macrophage colony-stimulating factor."; Kelvin et al. (2011), "Inflammatory cytokine expression is associated with chikungunya virus resolution and symptom severity."; Ng et al. (2009), "IL-1beta, IL-6, and RANTES as biomarkers of Chikungunya severity."; Teng et al. (2015).

[33] Burt et al. (2017).

[34] Miner et al. (2015).

Phenotype/covariate	Participants (N=42)
Gender	
Female, no. (%)	11 (26)
Male, no. (%)	31 (74)
Age	
Years, mean ± SD	9.22 ± 4.5
1-4 years old (%)	9 (21)
5-8 years old (%)	9 (21)
9-14 years old (%)	23 (57)
Signs or symptoms at enrollment	
Days post symptom onset, mean ± SD	1.41 ± 0.5
Fever, mean temperature ± SD, °C	38.3 ± 0.8
Fever, mean duration ± SD, days	2.4 ± 0.6
Peak fever >38.5°C (%)	16 (38)
Retroorbital pain (%)	7 (16)
Osteomuscular pain (%)	26 (62)
Rash (%)	41 (98)
Arthralgia (%)	36 (86)
Myalgia (%)	20 (48)
Headache (%)	9 (21)
Abdominal pain (%)	13 (31)
Vomiting (%)	4 (10)
Fluid accumulation (%)	14 (33)
Hospitalized (%)	36 (86)
Laboratory values at enrollment	
Median platelet count, mm^{-3} (range)	199,000 (88,000–337,000)
Nadir platelet count <100,000 mm^{-3} (%)	8 (19)
Median white cell count, mm^{-3} (range)	8,140 (3,030–16,120)
Median monocyte % of WBCs (range)	8.9 (4.1–14.6)
Median lymphocyte % of WBCs (range)	39.7 (10.4–66.4)
Laboratory values at convalescent timepoint	
Days post symptom onset, mean ± SD	15.7 ± 0.6
Median CHIKV IgG Ab titer, dilutions (range)	1,458 (232–7,794)
Severity categorization[a]	
Severe (%)	21 (50)
Non-severe (%)	21 (50)

Table 6.1: **Clinical characteristics of study population.** Abbreviations: WBC, white blood cell; CHIKV, chikungunya virus; IgG, immunoglobulin G; Ab, antibody.

[a] Cases were categorized as severe if the patient had either a peak fever of >38.5°C or a nadir platelet count of <100,000 mm^{-3}.

Results

Clinical characteristics of study participants

Blood samples for this study were collected from the Nicaraguan National Pediatric Reference Hospital in Managua, Nicaragua during routine care. Patients were sampled and tested following presentation with symptoms and histories suggestive of DENV or CHIKV infection, and a diagnosis of CHIKV viremia was confirmed by either RT-PCR or viral isolation assay. A total of 42 pediatric cases of CHIKV viremia presenting between November 2014 and October 2015 were included, from which acute (1-2d p.s.o.) and convalescent (15-17d p.s.o.) samples were obtained, for a total of 84 samples. De-identified clinical data were obtained for all samples and are summarized in Table 6.1. 74% of the patients were male, and over half were between 9-14 years of age. Rash was the most common presenting symptom (41/42, 98%), followed by arthralgia (36/42, 86%). Most of the patients were febrile (mean temperature, 38.3°C) and the average fever duration was 2.4 days. Of the cases studied, 36/42 (86%) resulted in hospitalization. We adapted a previous rubric for classifying CHIKV cases as non-severe vs. severe.[35] In this study, a severe case is defined by a peak temperature >38.5°C or a nadir platelet count <100,000 mm^{-3}. By this criterion, half of the pediatric cases (21/42, 50%) were considered severe.

[35] Chow et al. (2011); Ng et al. (2009).

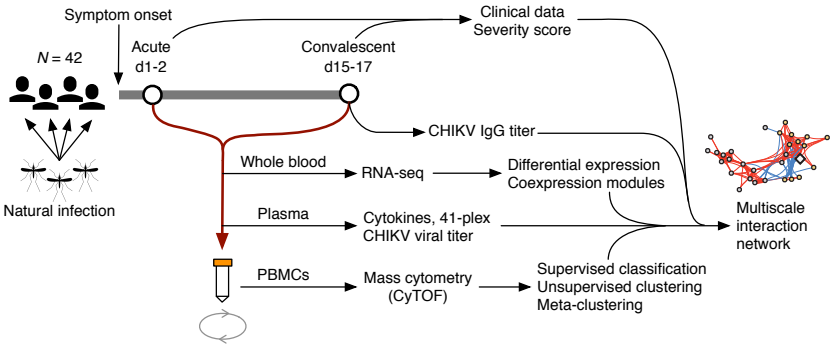


Figure 6.1: **Study design.** Blood samples were obtained from 42 pediatric cases of natural chikungunya (CHIKV) infections at an acute (d1-2) and a convalescent (d15-17) timepoint, relative to reported symptom onset. Samples were separated into whole blood, serum, and peripheral blood mononuclear cell (PBMC) aliquots for transcriptomic analysis, CHIKV viral titer assays, multiplex ELISA for cytokines, and mass cytometry (CyTOF), respectively. These data were combined with clinical data, including a severity score and a d15-17 CHIKV immunoglobulin G (IgG) titer, to create a multiscale network of interactions during the observed course of CHIKV infection.

Sampling times closely adhered to the targeted acute (SD = 0.5d) and convalescent (SD = 0.6d) timepoints. Each blood sample was separated into aliquots of whole blood, serum, and PBMCs for transcriptomic analysis via RNA-seq, CHIKV viral titer assays, multiplex ELISA for cytokines, and mass cytometry as illustrated in Figure 6.1. These data were then analyzed for changes correlat-

ing with the acute and convalescent phases of infection, severe and non-severe cases, the 15d immunoglobulin G (IgG) titer, and the acute phase viral titer. The resulting signatures and clusters were combined into a multiscale interaction network capturing the global landscape of immune responses to CHIKV.

Acute infection associates with CD14⁺CD16⁺ monocyte expansion

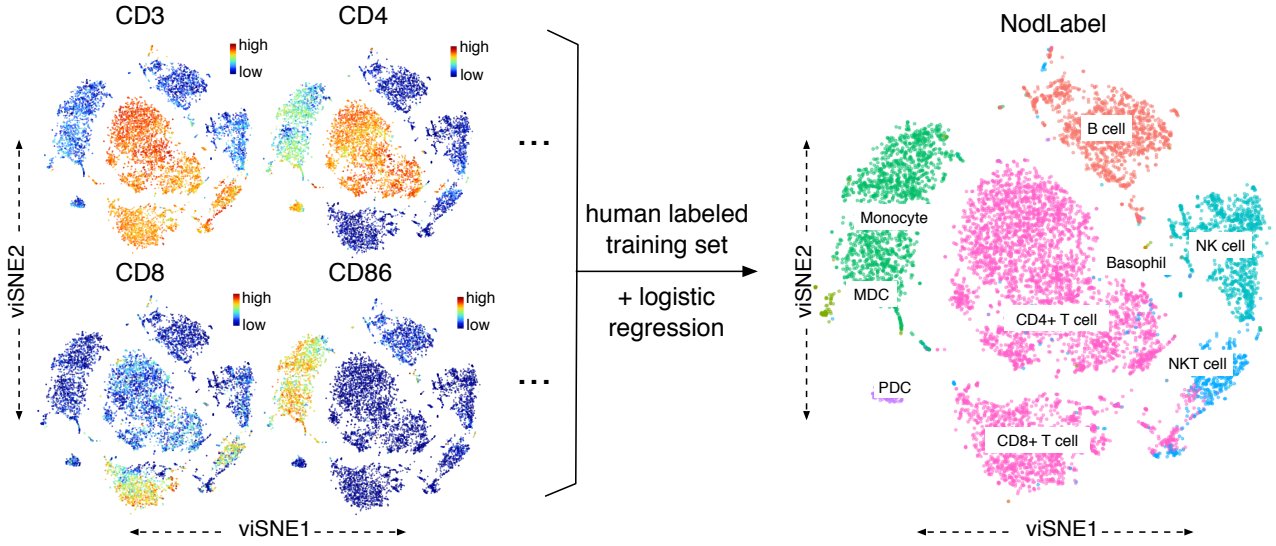


Figure 6.2: **Overview of the `NodLabel` procedure**, using a viSNE layout of CyTOF single-cell events. Left side, point color indicates channel values for four example channels. Right side, traditional hierarchical gating was used on a subset of samples to identify 9 major immune compartments, which was then used to train a logistic regression classifier (Nod) that applied labels for canonical leukocyte phenotypes to all samples (NodLabel).

CyTOF uses metal-labeled reagents and inductively coupled plasma mass spectrometry to overcome the limits of fluorescence spectral overlap in flow cytometry, allowing measurement of up to 50 analytes at single-cell resolution. We used CyTOF to quantify 35 immune cell surface markers (Appendix Table A.2) and the CHIKV surface protein in each of our samples. The high dimensionality of CyTOF data presents challenges for applying the traditional gating methods used in lower-dimensional flow cytometry. To address these challenges, we developed a sequential, semi-supervised approach to identify and classify immune cell populations in the CyTOF dataset. Manual gating and human-authored labels were first applied to a subset of the data to train a logistic regression classifier (called `NodLabel`) that was run on the remaining samples to broadly define 9 major immune subsets in each of the patient samples. Figure 6.2 illustrates this process using a viSNE layout algorithm[36] for 2D visualizations of high-dimensional CyTOF data from a representative patient sample.

[36] Amir et al. (2013), "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia."

To define additional heterogeneity within each of these broad subsets, we next applied Louvain/Phenograph[37] as an unsupervised clustering method to each `NodLabel`-identified subset in each patient sample. While the combination

[37] Levine et al. (2015), "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis".

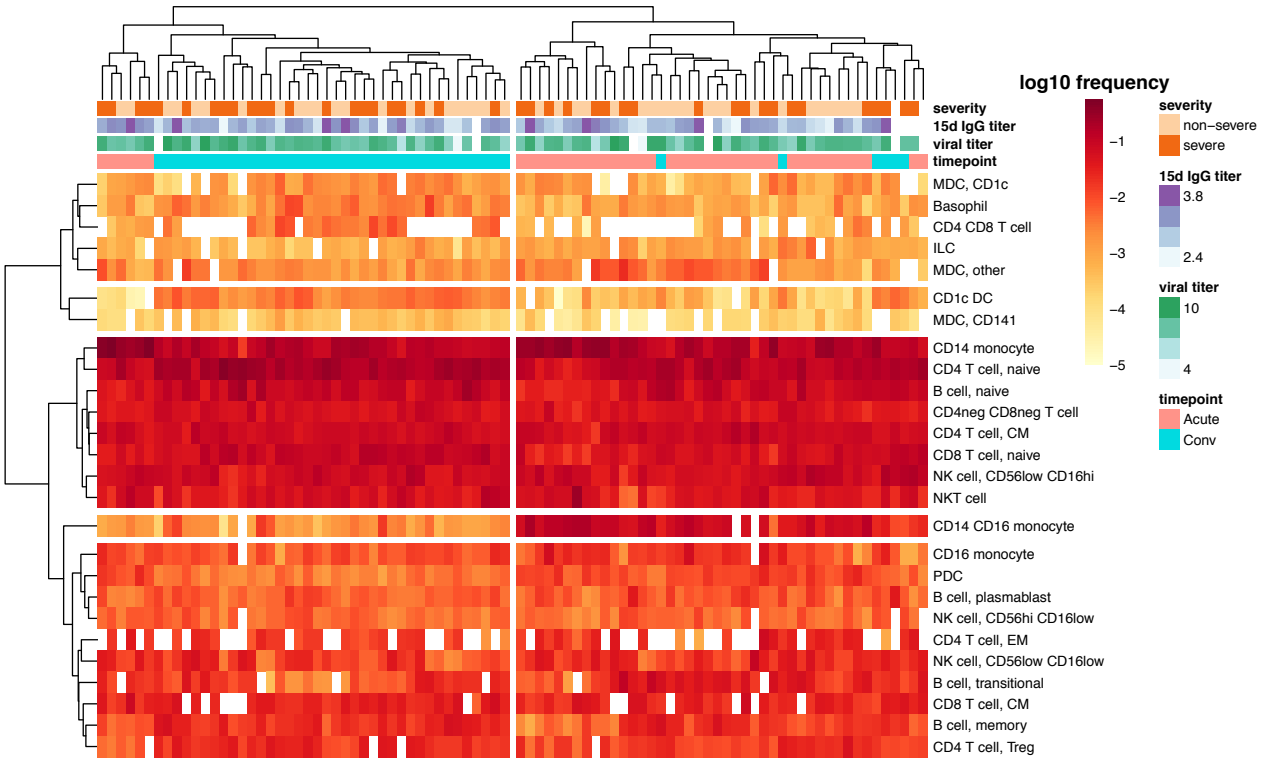


Figure 6.3: **CyTOF reveals signatures for acute CHIKV infection based on canonical immune cell phenotype clustering.** Heatmap of $\log_{10}$ scaled peripheral blood mononuclear cell (PBMC) community frequencies for all samples. Clinical variables are depicted for all samples across the top of the heatmap; 15d post symptom onset immunoglobulin G (IgG) titer and viral titer (which was measured during the acute phase) are both in units of $\log_{10}$ dilutions. Hierarchical clustering (using complete linkage) was applied to both samples (X axis) and communities (Y axis). Four major clusters of communities and two major clusters of samples (largely separating acute and convalescent samples) are highlighted.

of the `NodLabel` classifier and Phenograph (which we term `HybridLouvain`) is a powerful approach to define phenotypic heterogeneity in a sample, a limitation of the approach is that the identified `HybridLouvain` communities are only applicable to a single patient sample. To address this issue, we meta-clustered the individual communities across all patient samples, and meta-communities that were reproducibly identified across multiple patients were then manually annotated based on overall marker expression patterns. These annotations were then mapped back to each individual patient sample to provide consistent community definitions across all samples. Using this approach (which we term `MetaHybridLouvain`) we identified 26 communities of canonical leukocyte populations across all acute and convalescent phase samples (Figure 6.3). Hierarchical clustering by sample revealed two clusters that generally separated by timepoint, with no communities corresponding to any apparent contrasts in severity,

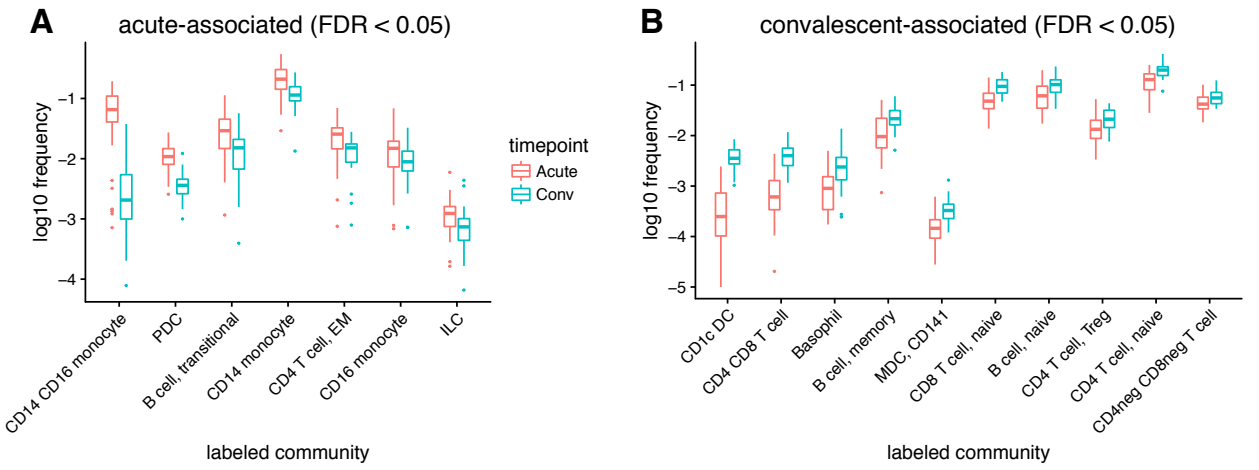


Figure 6.4: **PBMC communities with differing frequency across the timepoints of CHIKV infection.** A, $\log_{10}$ frequencies for PBMC communities contrasted between acute and convalescent phase samples, filtered to communities where the acute phase frequency was higher at a significance threshold of FDR < 0.05 (Mann-Whitney *U*). B, same as A but filtered to communities where the convalescent phase frequency was higher at a significance threshold of FDR < 0.05.

15d IgG titer, or acute phase viral titer (Figure 6.3). Clustering by community frequency reveals a distinct contrast in $CD14^+CD16^+$ monocyte frequencies between the acute and convalescent phases (vertical axis, Figure 6.3). This is the most expanded population at the acute timepoint (Figure 6.4A), and the difference was highly significant (Bonferroni-corrected [BF] P = 1.9e-09). Other populations also comparatively upregulated during the acute phase were plasmacytoid dendritic cells (PDCs) (BF P = 4.7e-13) and $CD14^+$ monocytes (BF P = 0.0019), with four additional populations identified at a threshold false discovery rate (FDR) < 0.05 (Figure 6.4A). Ten populations were comparatively downregulated at the acute phase and thereby convalescent-associated at FDR < 0.05 (Figure 6.4B), with the strongest difference being observed in CD1c dendritic cells (DCs) (BF P = 3.9e-19).

Monocytes, dendritic cells, and B cells express CHIKV surface protein during acute infection

Classifying CyTOF events into only the canonical leukocyte populations ignores much of the richness of these data, which can reveal previously unrecognized diversity and heterogeneity within each of these populations. The advantage of our `MetaHybridLouvain` approach (Figure 6.5) is that it allows for unbiased identification of phenotypically heterogenous sub-communities within each of the canonical immune subsets,[38] e.g., a sub-community of $CD14^+$ monocytes as defined by a specific, reproducible marker expression

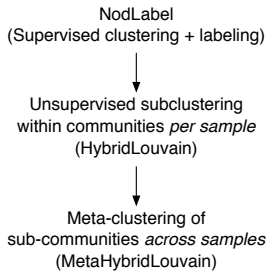


Figure 6.5: Overview of the `MetaHybridLouvain` procedure.

[38] As opposed to strict classification methods, e.g. Lee et al. (2017), which require predefining all desired cell types. A complete comparison of CyTOF classification and clustering methods is beyond this chapter's scope, but Aghaeepour et al. (2013) provides an introduction, and Samusik et al. (2016) is a comparable clustering method that also emphasizes novel subpopulation discovery.

pattern across multiple samples. This allowed for the identification of up to nine sub-communities within certain defined canonical immune populations (Figure 6.6, 6.7), producing a total of 57 sub-communities. More detailed results of `MetaHybridLouvain` for the representative sample used in Figure 6.2 and 6.6 are depicted in Appendix Figures B.1-B.2.

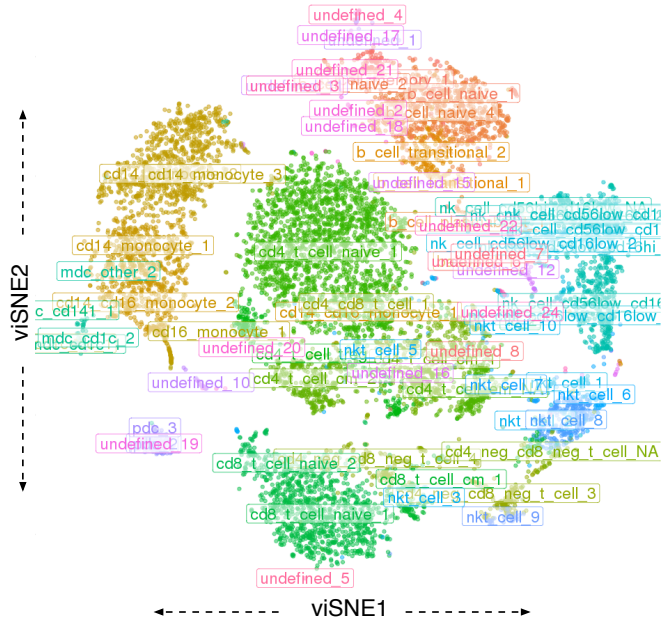


Figure 6.6: viSNE layout of CyTOF single-cell events from the same representative sample as Figure 6.2, now with labels for peripheral blood mononuclear cell (PBMC) sub-communities detected by `MetaHybridLouvain` drawn at the centroid of each sub-community.

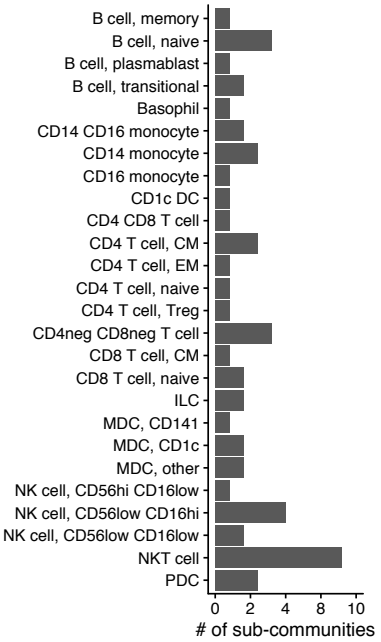


Figure 6.7: Number of sub-communities detected by `MetaHybridLouvain` for each of the canonical leukocyte phenotypes.

Having dissected the cellular heterogeneity of the samples at high resolution, we went on to identify leukocyte populations that have comparatively high levels of CHIKV surface protein during the acute phase of infection. Qualitatively, in the representative patient sample, CHIKV surface protein was expressed by distinct populations of leukocytes in the viSNE layout (Figure 6.8)—in particular monocytes, dendritic cells, and B cells (compare with Figures 6.2 and 6.6). Quantitatively, across all samples, monocytes and dendritic cells displayed the strongest contrasts in mean CHIKV channel values per sample between acute and convalescent timepoints (Figure 6.9). CHIKV-positive cell populations largely fell along canonical leukocyte phenotype boundaries, with all three sub-communities of $CD14^+$ monocytes (BF P = 2.5e-26, 2.7e-24, 8.1e-16), both sub-communities of CD1c MDCs (BF P = 4.2e-15 and 4.3e-08), all CD1c DCs (BF P = 4.3e-16), and both sub-communities of $CD14^+CD16^+$ monocytes (BF P = 2.5e-09 and 2.3e-06) identified as significantly more CHIKV-positive during the acute phase. Interestingly, although the differences were less pronounced,

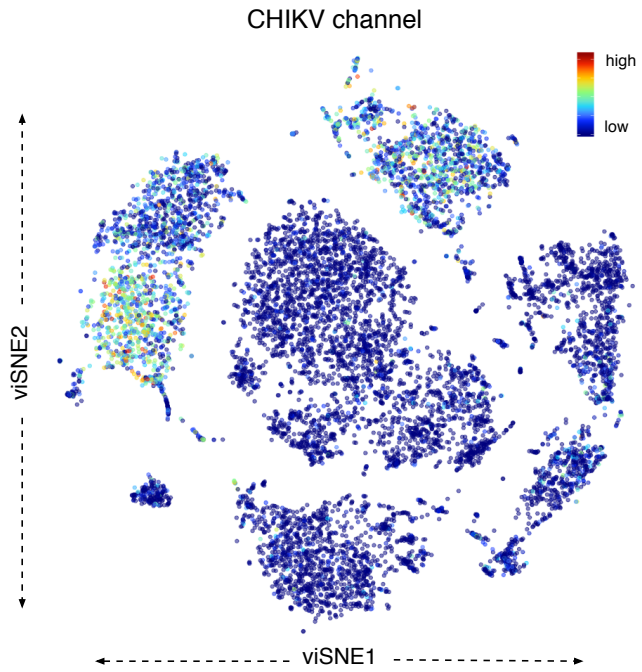
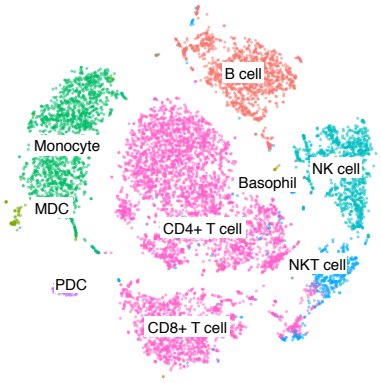


Figure 6.8: viSNE layout of CyTOF single-cell events from the same representative sample as Figure 6.2 and 6.6, now with points colored according to the CHIKV channel. By qualitative comparison with Figure 6.2 (reproduced below), monocytes, myeloid dendritic cells (MDCs), and B cells have the highest CHIKV surface protein expression.



three sub-communities of B cells were also observed to express significantly higher CHIKV surface protein during acute infection: the only community of memory B cells (BF P = 3.2e-09) and two of the four sub-communities of naïve B cells (BF P = 9.7e-06 and 0.00022). Although CHIKV surface protein expression only correlates with (and does not establish) tropism of the virus, our data suggest that among PBMCs, CHIKV preferentially infects monocytes and dendritic cells, while displaying lower but substantial affinity for B cells.

CD14$^+$ and CD14$^+$CD16$^+$ monocyte sub-communities exhibit contrasting behaviors during acute infection

Although CD14$^+$CD16$^+$ monocytes expand during the acute phase of infection, community subclustering provides more detail on particular sub-communities that associate with the acute phase. Hierarchical clustering of samples by sub-community frequencies separates the samples by timepoint more effectively than canonical population frequencies alone (Figure 6.10, compare with Figure 6.3), with only 3/88 (3%) of samples misclassified between the two major clusters. Again, however, there was no apparent clustering of samples that corresponded to contrasts in clinical severity, 15d IgG titer, or acute phase viral titer. Stratifying by the acute and the convalescent phases, there were no significant

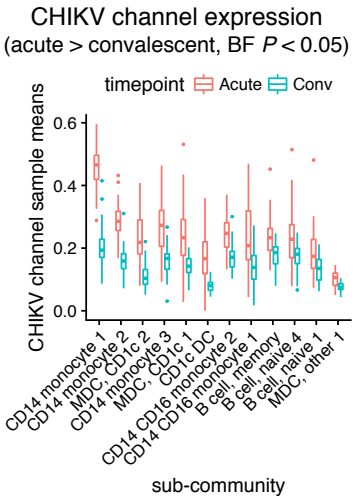


Figure 6.9: Differences in CHIKV surface protein expression between acute phase and convalescent phase samples per `MetaHybridLouvain` sub-community. Sub-communities are ordered by largest to smallest difference and filtered to sub-communities where the median of the channel means per sample was higher in the acute phase samples at a significance threshold of Bonferroni $P < 0.05$. Sub-communities are named by their parent canonical community name plus an arbitrary number, up to the counts given in Figure 6.7.

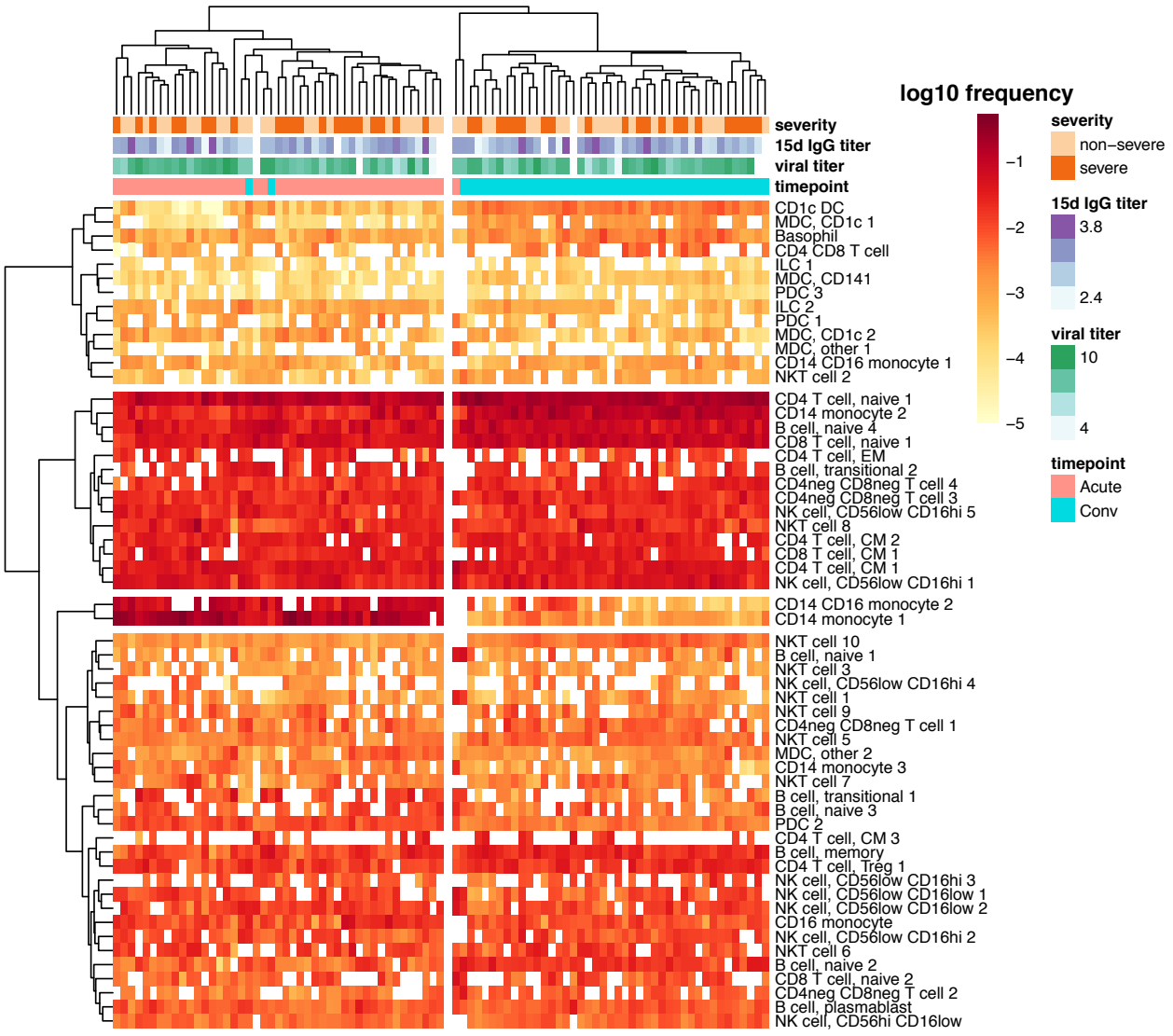


Figure 6.10: **Specific monocyte sub-communities undergo expansion during acute CHIKV infection.** Heatmap of $\log_{10}$ scaled PBMC sub-community frequencies for all samples. Clinical variables are depicted for all samples across the top of the heatmap; 15d post symptom onset immunoglobulin G (IgG) titer and viral titer (which was measured during the acute phase) are both in units of $\log_{10}$ dilutions. Hierarchical clustering (using complete linkage) was applied to both samples (X axis) and sub-communities (Y axis). Four major clusters of sub-communities and two major clusters of samples (largely separating acute and convalescent samples) are highlighted.

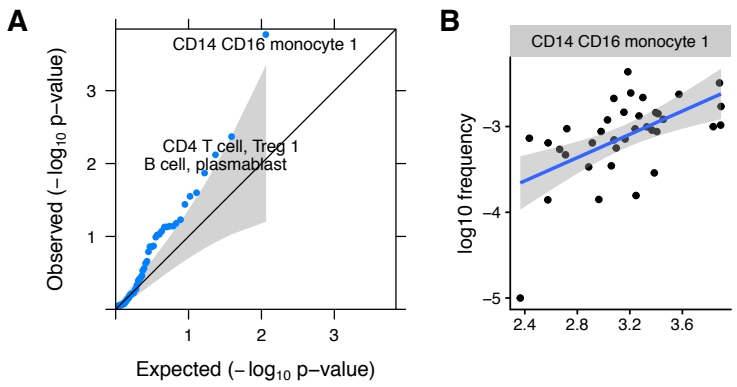


Figure 6.11: **Correlations between acute phase cell sub-community frequencies and $\log_{10}$ CHIKV IgG titer at 15d post-symptom onset.** A, Q-Q plot of the distribution of observed $-\log_{10} P$ values against the distribution of $-\log_{10} P$ values expected under the null hypothesis (that Spearman's $\rho = 0$ for all sub-communities). Gray shaded band indicates the 95% confidence interval for the expected P value distribution under the null hypothesis. B, Scatterplots of $\log_{10}$ cell sub-community frequencies against $\log_{10}$ IgG CHIKV titer at the 15d timepoint for the $CD14^+CD16^+$ monocyte 1 correlation, which is the only correlation significant after multiple hypothesis correction (BF $P = 0.0097$, Spearman's $\rho = 0.60$).

differences in any sub-community frequencies between severe and non-severe cases at FDR < 0.1. Within either timepoint, there were also no significant correlations between sub-community frequencies and log-transformed acute phase viral titers at FDR < 0.1. There was, however, a single significant correlation between $CD14^+CD16^+$ monocyte sub-community 1 at the acute phase and the 15d (convalescent) IgG titer (BF $P = 0.0097$, Spearman's $\rho = 0.60$; see Figure 6.11), with no significant correlations among convalescent sub-community frequencies at FDR < 0.1.

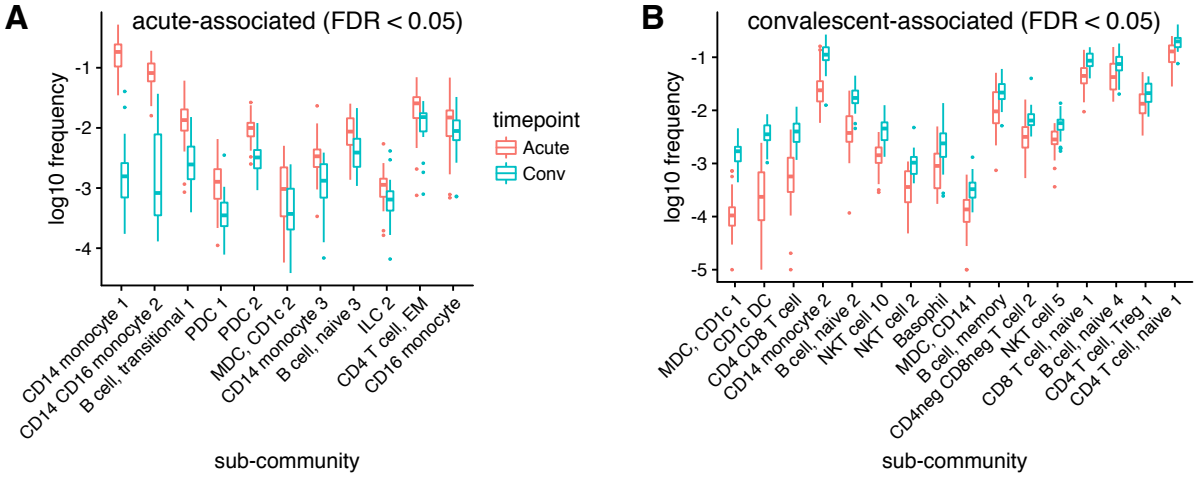


Figure 6.12: **PBMC sub-communities with differing frequency across the timepoints of CHIKV infection.** A, $\log_{10}$ frequencies for PBMC sub-communities contrasted between acute and convalescent phase samples, filtered to sub-communities where the acute phase frequency was higher at a significance threshold of FDR < 0.05 (Mann-Whitney U). B, same as A but filtered to sub-communities where the convalescent phase frequency was higher at a significance threshold of FDR < 0.05.

Among sub-communities significantly expanded during the acute phase, two particular expansions separated from the others by an order of magnitude (Figure 6.12A), specifically sub-community 1 of $CD14^+$ monocytes (BF P = 7.7e-21) and sub-community 2 of $CD14^+CD16^+$ monocytes (BF P = 2.8e-15). When examining the other two sub-communities of $CD14^+$ monocytes, one is

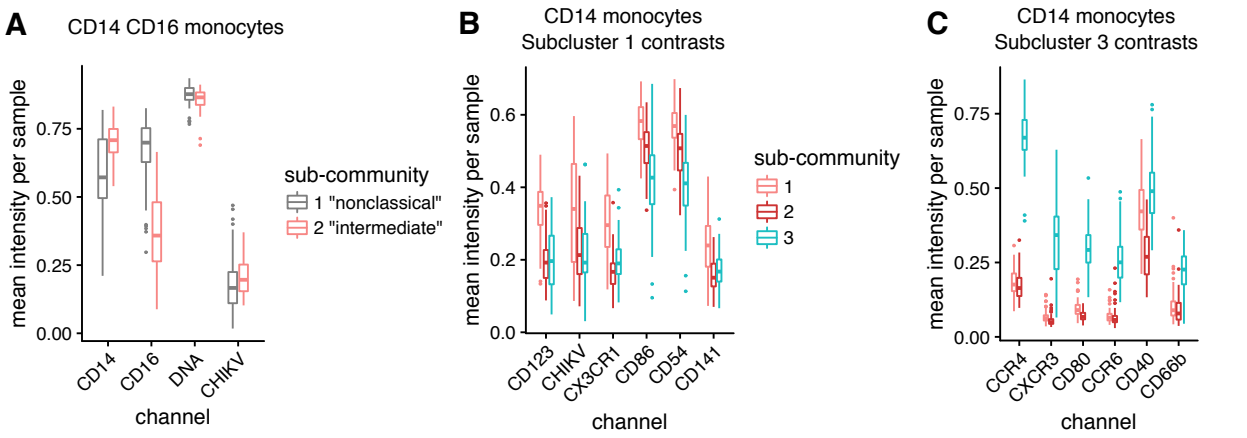


Figure 6.13: **Marker expression differences between sub-communities of $CD14^+CD16^+$ monocytes and $CD14^+$ monocytes**, depicted as boxplots of the mean expression levels for all samples. A, relative expression of $CD14^+$ and $CD16^+$ in $CD14^+CD16^+$ sub-communities indicates that sub-community 1 is a $CD14^+CD16^{++}$ (aka "non-classical") phenotype, while sub-community 2 is a $CD14^{++}CD16^+$ (aka "intermediate") phenotype. Differences shown here are significant at FDR < 0.05; for a view of all differences significant at this threshold, see Appendix Figure B.3. B, relative expression of six markers that most differentiate (by the difference in medians) sub-community 1 of $CD14^+$ monocytes from the other sub-communities. C, relative expression of six markers that most differentiate (by the difference in medians) sub-community 3 of $CD14^+$ monocytes from the other sub-communities. Note: Channels shown in B and C are a subset of the differences that are significant at FDR < 0.05; for a view of all differences significant at FDR < 0.05 see Appendix Figure B.6.

also expanded during the acute phase but to a lesser extent (sub-community 3, BF P = 5.6e-06) while the other instead is expanded during the convalescent phase (sub-community 2, BF P = 1.6e-10). At FDR < 0.1, sub-community 1 of $CD14^+CD16^+$ monocytes, which is the only other sub-community of $CD14^+$ $CD16^+$ monocytes, is not significantly different across timepoints. Other sub-communities associating with the convalescent phase at FDR < 0.05 include MDCs, CD1c DCs, B cells, T cells, and basophils (Figure 6.12B).

Since different sub-communities of $CD14^+CD16^+$ and $CD14^+$ monocytes displayed distinctive responses to acute CHIKV infection, we looked for marker differences that could better define these sub-communities. Examination of the two $CD14^+CD16^+$ monocyte sub-communities (Figure 6.13A) revealed that among all significant marker differences, sub-community 1 had higher CD16 expression (BF P = 5.7e-19) and sub-community 2 had higher CD14 expression (BF P = 6.4e-05). This corresponded to sub-communities commonly called "nonclassical" $CD14^+CD16^{++}$ and "intermediate" $CD14^{++}CD16^+$ monocytes,[39] implying that in our study, "intermediate" monocytes were substantially expanded during acute CHIK infection while "nonclassical" monocytes were unchanged. Significant contrasts in the expression of many other surface markers at FDR < 0.05 (Appendix Figure B.3) and the consistent identification of these patterns across the majority of samples (Appendix Figures B.4-B.5) confirmed the distinction between these sub-communities.

Among the three sub-communities of $CD14^+$ monocytes, we discovered two that were associated with acute infection, including one with a previously unreported phenotype. sub-community 1 (the sub-community most strongly as-

[39] Wong et al. (2011), "Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets."; Ziegler-Heitbrock et al. (2010), "Nomenclature of monocytes and dendritic cells in blood".

sociated with acute infection and also expressing the highest levels of CHIKV surface protein) was characterized by having relatively higher levels of CD123, CX3CR1, CD86 and CD54 expression (Fig 6.13B), generally consistent with a more activated phenotype relative to sub-community 2, which was more prevalent during convalescence. Monocyte sub-community 3 was also expanded during acute infection, though at a much lower frequency than monocyte sub-community 1, and displayed similar levels of CD40, consistent with an activated phenotype. Interestingly, however, this subset also exhibited comparatively high expression of markers that are not classically associated with monocytes, particularly the chemokine receptor CCR4, as well as CXCR3 and CCR6 (Figure 6.13C). We further confirmed that this sub-community did not express canonical markers associated with other major cell types, such as T cells or B cells, to verify that it did not represent an artifact of cell-cell doublets. Again, significant contrasts in the expression of many surface markers at FDR < 0.05 (Kruskal-Wallis test, Appendix Figure B.6), and a consistent pattern for the phenotype identified across the majority of samples (Appendix Figures B.7-B.9) confirmed the distinction between these sub-communities. Given the strongly contrasting associations of these sub-communities with the phase of infection, these data suggest that unappreciated heterogenity within $CD14^{+}$ monocyte phenotypes may enable different roles for sub-communities of these monocytes during CHIKV infection.

Monocyte-associated cytokine concentrations increase during acute infection

To profile the effect of CHIKV on circulatory markers for inflammation and immune signaling, we used a multiplexed microbead immunoassay (Luminex) to measure serum concentrations of 41 cytokines, chemokines, and growth factors in all 84 samples. In our study, seven cytokines were significantly different (Wilcoxon signed-rank test with Benjamini-Hochberg adjustment) across acute and convalescent timepoints at FDR < 0.05 (Figure 6.14A-E). The strongest contrast was the IFN-γ inducible, monocyte-secreted chemokine CXCL10 (BF P = 0.00085). Significant increases (FDR < 0.05) were also observed for IL-10 (a monocyte-secreted anti-inflammatory cytokine), CCL2 (monocyte chemoattractant protein 1), IFN-α2, TNF-α, IL-8, and IL-1RA. There were no significant differences observed among growth factors or colony-stimulating factors (Ap-

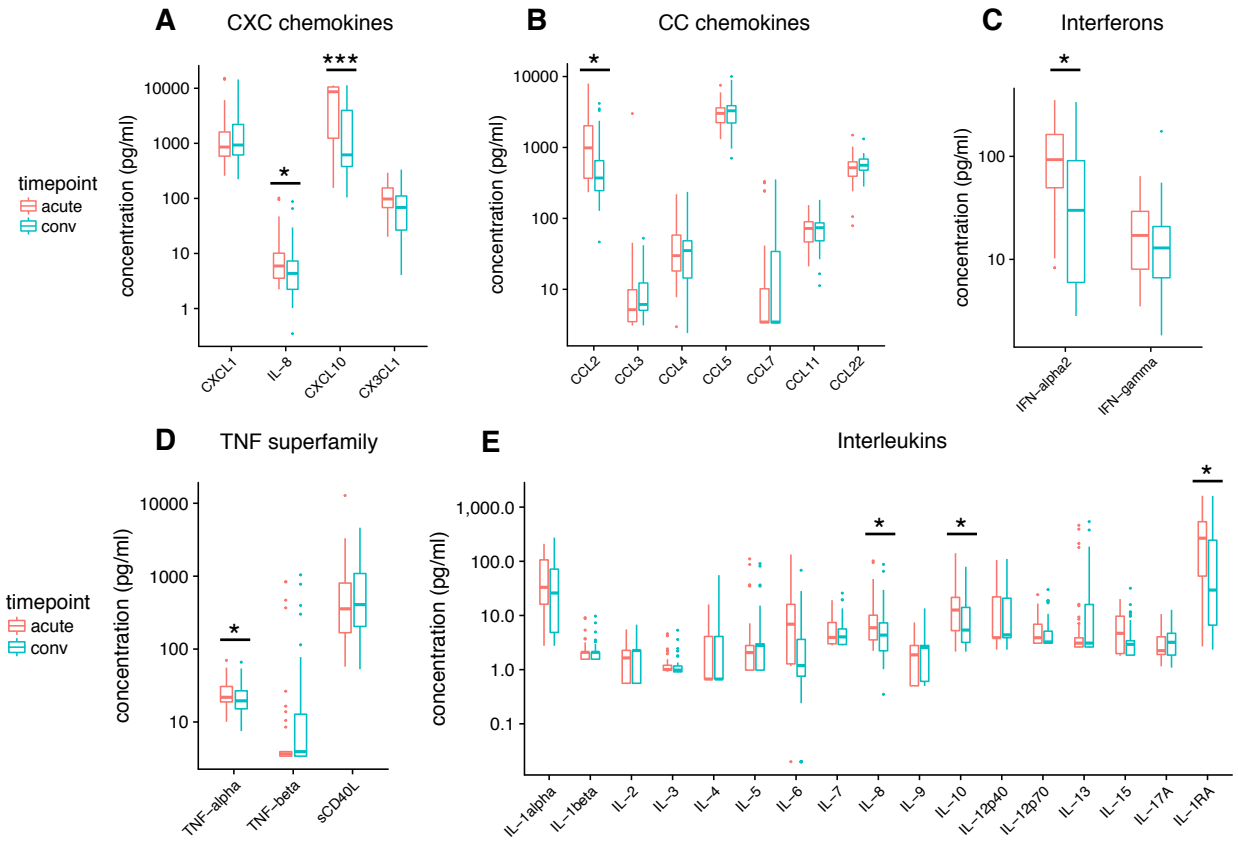


Figure 6.14: **Differences in serum cytokine and chemokine levels between the acute and convalescent phase samples.** Wilcoxon signed-rank test was used to determine statistical significance, with a Benjamini-Hochberg adjustment for FDR (n = 41). $*P < 0.05$, $***P < 0.001$. A, CXC chemokines. B, CC chemokines. C, interferons. D, TNF superfamily cytokines. E, interleukins. Note that IL-8 is both a CXC chemokine (CXCL8) and an interleukin, so it is depicted twice. Growth factors and colony-stimulating factors are shown in Appendix Figure B.10.

pendix Figure B.10). No analyte concentrations were significantly decreased during acute infection.

To determine if cytokine levels could be associated with changes in specific cell sub-communities, we correlated log-scaled Luminex analyte concentrations with log-scaled sub-community frequencies (using Pearson's r). Hierarchical clustering revealed that cytokine and growth factors concentrations tended to correlate with each other rather than with any of the subpopulation frequencies (Appendix Figure B.11). This remained unchanged when stratifying into acute phase (Appendix Figure B.12) or convalescent phase (Appendix Figure B.13) samples, with the only exception being a cluster containing CXCL1, sCD40L, PDGF-AA, and PDGF-AB/BB that consistently separated from one major cluster containing all other Luminex analytes. Since the most pronounced expansions during the acute phase involved monocyte subpopulations, we then performed a more focused analysis on correlations between all

cytokines and monocyte subpopulations, stratifying by timepoint to capture potential regulatory relationships rather than the primary contrast of the study. Within acute phase samples, cytokines generally had varied correlations with each of the monocyte subpopulations, but at the convalescent timepoint, the monocyte chemoattractant CCL2 clustered separately from all other cytokines and had positive correlations with all monocyte subpopulations (Figure 6.15, BF $P = 0.0021$). This is suggestive of a relatively important regulatory role for CCL2 on monocyte populations during the convalescent phase of infection.

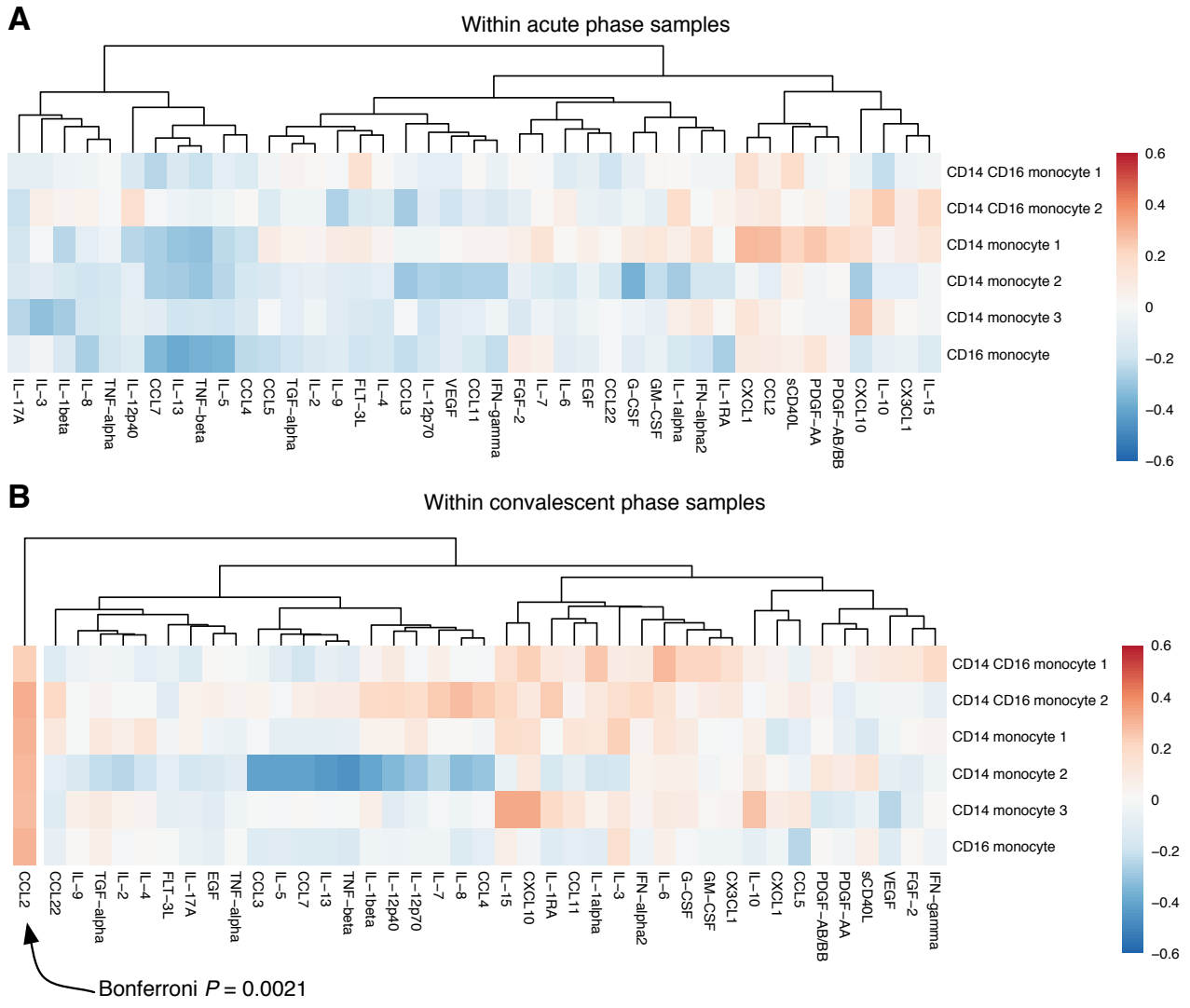


Figure 6.15: **Clustered heatmap of Pearson correlations between log-scaled serum cytokine concentration and log-scaled monocyte subphenotype frequencies.** A, within acute phase samples. B, within convalescent phase samples.

Acute infection associates with upregulated transcription of monocyte-associated cytokine genes

To capture global transcriptional changes during CHIKV infection, polyadenylated RNA from whole blood was analyzed by RNA-seq for all 42 patients at both sampled timepoints with two technical replicates per sample. A Tuxedo pipeline was used for read alignment, quantification and differential expression analysis of genes. Since previous studies of gene and protein changes during acute CHIKV infection targeted cytokines, chemokines, and innate immunity mechanisms,[40] we first present results for differentially expressed genes in these pathways for comparison with our Luminex data and the literature, before moving to a global analysis in the subsequent section.

[40] Chow et al. (2011); Hoarau et al. (2010), "Persistent chronic inflammation and infection by Chikungunya arthritogenic alphavirus in spite of a robust host immune response."; Ng et al. (2009); Teng et al. (2015); Wauquier et al. (2011), "The acute phase of Chikungunya virus infection in humans is associated with strong innate immunity and T CD8 cell activation".

Appendix Figure B.14 shows that across the two timepoints, log-scaled serum concentrations for the significantly CHIK-upregulated cytokines (as measured by Luminex, Figure 6.14) did not correlate with whole blood gene expression for corresponding genes, using log-scaled units of fragments per kilobase of exon per million reads mapped (FPKM). Since this could be due to different "baseline" serum cytokine or cytokine gene expression levels, we repeated the analysis with all acute phase measurements normalized against the convalescent phase measurements, but clustering did not change (Appendix Figure B.15). This suggests that the regulation of serum cytokine levels is not primarily driven by transcriptional changes in leukocytes, but could involve substantial expression from other tissues and secretory and protein-level regulatory processes. Alternatively, the technical variance of the Luminex assay for our samples may simply have been too high to draw out meaningful correlations with gene expression.

Differential expression of all genes was quantified by $\log_2$ fold change (log2FC) in units of FPKM and considered significant at an FDR threshold of <0.05. Among CXC and CC subfamily chemokines, we observed transcriptional upregulation of *CXCL10*, *CXCL11*, *CCL2*, *CCL7*, and *CCL8* (Figure 6.16). Of these, monocyte-secreted *CXCL10* and monocyte chemoattractant *CCL2* concur with the changes in serum cytokine concentrations described above (Figure 6.14A-B), and *CCL8* (whose gene product was not measured with Luminex) is notable for being another monocyte chemoattractant (MCP-2). Interestingly, although serum IFN-α levels were significantly elevated during acute

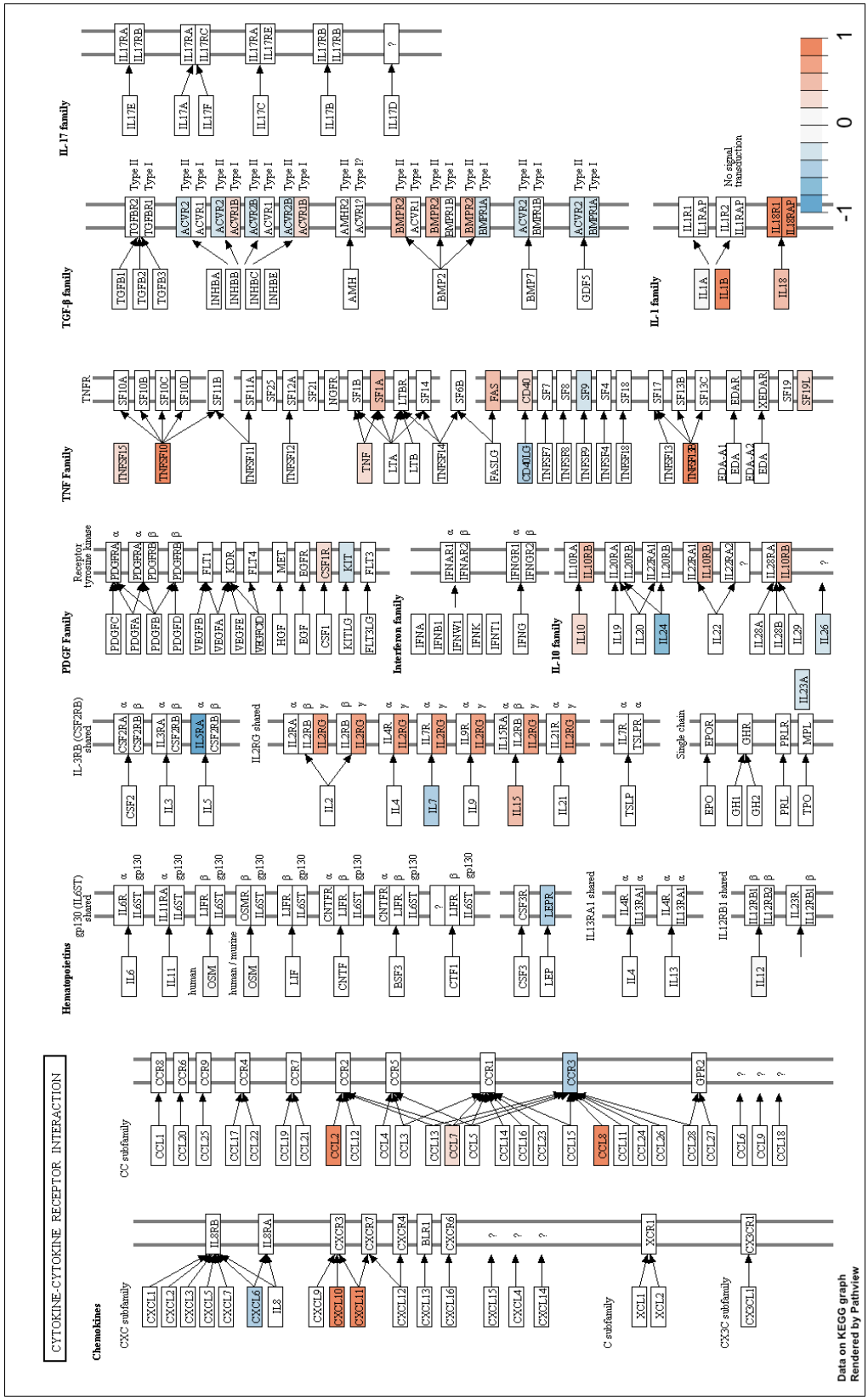


Figure 6.16: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the cytokine-cytokine receptor interaction pathway**, using KEGG annotations (accession hsa04060). Positive values indicate upregulation during the acute phase of infection.

infection (Figure 6.14C), none of the interferon family genes were differentially expressed (Figure 6.16). Upregulation of the *TNF* gene is concordant with the significant increase in serum TNF-α concentration (Figures 6.16 and 6.14D), and other significantly upregulated TNF superfamily genes included *TNFSF15*, *TNFSF10*, and *TNFSF13B* (Figure 6.16). While upregulation of *IL10* gene expression was concordant with the serum cytokine measurements, in contrast to those data, we did not observe differential expression of *IL8* (Figures 6.16 and 6.14E).

We also examined differential expression of genes during acute infection for known components of innate immunity pathways annotated in KEGG[41] (Appendix Figures B.16-B.21). Of these pathways, JAK-STAT signaling genes were significantly transcriptionally upregulated, with smaller but significant upregulation of NF-κB genes (Appendix Figures B.16 and B.19). Among toll-like receptor (TLR) genes, *TLR5* and *TLR3* (whose gene product is a dsRNA receptor) were significantly upregulated (Appendix Figure B.17). Both *RIG-I* and *MDA5*, which are cellular sensors for viral RNA, were significantly upregulated (Appendix Figure B.18). Of the TNF superfamily receptors, *TNFR1* was upregulated but *TNFR2* was not (Appendix Figure B.20). Of the interferon regulatory factor (IRF) genes annotated in KEGG, expression of *IRF7* and *IRF9* were significantly upregulated, and interestingly, all downstream transcriptional targets of IRF9 annotated in KEGG (*MX1*, OAS genes, *ADAR*, and *PML*) were consistently upregulated during acute infection (Appendix Figure B.21). (Fold change values for all quantified genes and *q* values are provided in Supplementary Data S2.) In general, our observed modulations of human innate immunity pathways were comparable to differentially expressed genes reported for a mouse model in a recent study[42] (see Discussion).

[41] Ogata et al. (1999), "KEGG: Kyoto Encyclopedia of Genes and Genomes".

[42] Wilson et al. (2017), "RNA-Seq analysis of chikungunya virus infection and identification of granzyme A as a major promoter of arthritic inflammation."

[43] Anders, Reyes, and Huber (2012), "Detecting differential usage of exons from RNA-seq data."; Trapnell et al. (2013), "Differential analysis of gene regulation at transcript resolution with RNA-seq."

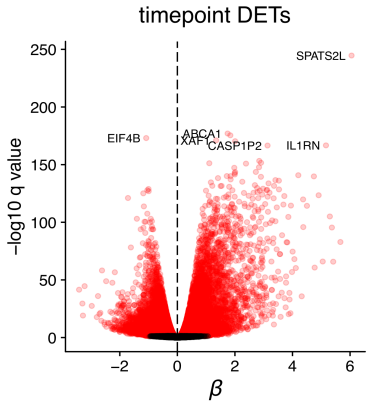


Figure 6.17: **Volcano plot of differentially expressed host transcripts between acute and convalescent phase samples**, with negative $\log_{10}$ scaled q values (Benjamini-Hochberg adjusted P values) on the Y axis, and the modeled β coefficient for each transcript (corresponding to natural log transformed effect sizes) on the X axis. Transcripts to the right of the vertical dashed line were comparatively upregulated in acute phase samples, while transcripts to the left were upregulated during the convalescent phase. Transcripts that pass FDR < 0.05 are colored red. Top transcripts by q value are individually labeled by their corresponding gene symbol.

Transcriptomic signatures for acute infection, severity, viral titer, and immunogenicity

RNA-seq enables the estimation of transcript abundances and transcript-level differential expression analyses that may offer insights not available from gene-level quantification.[43] We performed pseudoalignment-based quantification of transcript abundances in units of transcripts per million (TPM), followed by differential expression analysis.

After adjusting for age and gender covariates, a strong transcriptional signa-

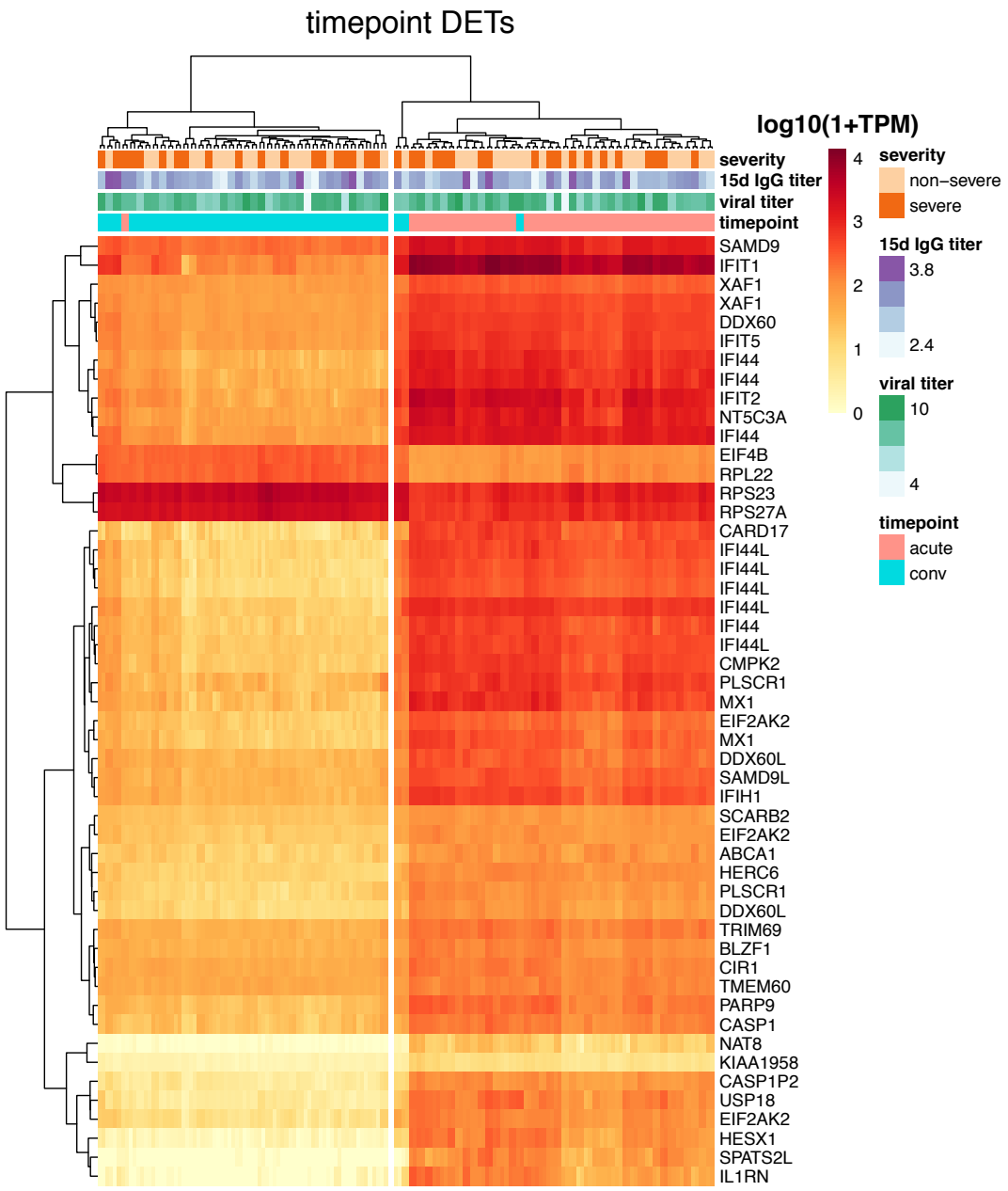


Figure 6.18: **Top 50 differentially expressed host transcripts for CHIKV infection phase.** Heatmap of expression measured in $\log_{10}$ scaled 1 + transcripts per million (TPM) for top 50 differentially expressed transcripts between acute and convalescent phase samples (2 technical replicates per patient sample). Clinical variables are depicted for all samples across the top of the heatmap; 15d post symptom onset immunoglobulin G (IgG) titer and viral titer (which was measured during the acute phase) are both in units of $\log_{10}$ dilutions. Hierarchical clustering (using complete linkage) was applied to both samples (X axis) and transcripts (Y axis). Two major clusters of samples (largely separating acute and convalescent samples) are highlighted.

ture for timepoint (acute vs. convalescent) emerged (Figure 6.17), with 28,015 transcripts differentially expressed at FDR < 0.05. The top differentially expressed transcripts (DETs), ordered by *P* value, were products of the *SPATS2L*, *IL1RN*, *EIF4B*, *ABCA1*, *XAF1*, and *CASP1P2* genes. Hierarchical clustering of the samples by quantification of the top 50 differentially expressed transcripts readily separated samples by timepoint, with only 8/160 (5%) of samples misclassified between the two major clusters (Figure 6.18). All paired technical replicates clustered together. The top 50 DETs notably contained many interferon-induced (IFI prefix) gene products that cluster along the vertical axis (Figure 6.18).

Adding the log-scaled acute phase CHIK viral titer as a variable to the model of transcript expression produced a separate and substantial signature for transcription correlating with viral titer (Figure 6.19), with 3,326 DETs at FDR < 0.05. In Figure 6.19, an increase in transcription corresponding to higher viral titers were modeled as a positive fixed effect coefficient β. Top DETs for viral titer (by *P* value) were products of the *PPT1*, *DDX52*, *LILRB3*, *CDS2*, and *FBXO7* genes, with all but the last upregulated in cases with higher viremia.

We assessed the composition of these two large signatures with standard gene set enrichment analyses for functional annotation libraries. The top 1,000 genes in the timepoint DET signature were most significantly enriched for gene ontology (GO), Panther, and Reactome annotations related to defense against viral infection, TLR signaling, and interferon signaling (Table 6.2). The top 1,000 genes in the viral titer DET signature were most significantly enriched for terms related to leukocyte activation, chemokine and cytokine signaling, and interferon signaling (Table 6.2). These enrichments suggest that the timepoint of infection sensibly corresponds to a maximal contrast in transcription of innate antiviral immunity genes, while the level of viremia correlates with increased transcription of immune cell activation and recruitment signals.

Although the phase of infection and the level of viremia were expected to produce strong transcriptional signatures, we sought potential signatures for downstream clinical outcomes, such as the severity of acute phase symptoms or the CHIKV IgG titer measured at 15d p.s.o., a correlate for humoral immunogenicity. For symptom severity, adding the severe vs. non-severe categorization of cases to the model produced a small differential expression signature of 56 transcripts at FDR < 0.05 (Figure 6.20), with *P* values for the top three tran-

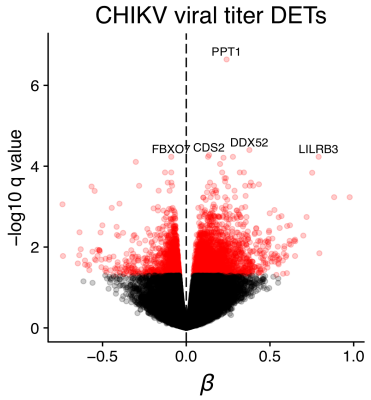


Figure 6.19: **Volcano plot** as in Figure 6.17 but for DETs between samples with higher and lower CHIKV viral titer. Transcripts to the right of the vertical dashed line associated with higher viral titer, while transcripts to the left associated with lower viral titer.

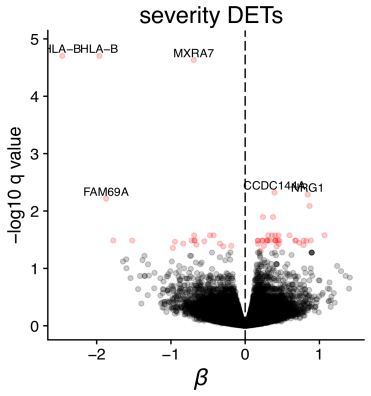


Figure 6.20: **Volcano plot** as in Figure 6.17 but for DETs between patients with more severe and less severe acute phase symptoms. Transcripts to the right of the vertical dashed line were comparatively upregulated in severe cases, while transcripts to the left were upregulated in non-severe cases.

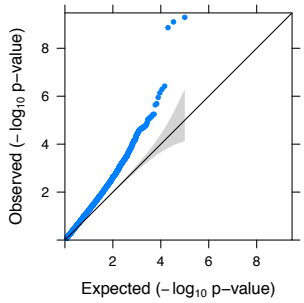


Figure 6.21: **Q-Q plot of the distribution of observed $-\log_{10}$ *P* values for severity DETs against the distribution expected under the null hypothesis.** Gray shaded band indicates the 95% confidence interval for the null distribution.

Gene set[a] (# genes)	Annotation set	Term	Overlap	q value[b]	Z-score	Combined score[c]
Top 1000 DETs for timepoint: acute vs. convalescent (593)	GO Biological Process 2015	defense response to virus	40/147	5.98e-24	-2.26	120
		response to virus	47/250	1.50e-21	-2.37	114
		viral life cycle	36/118	1.51e-23	-2.14	112
	Panther 2016	toll receptor signaling pathway	7/49	0.0425	-1.51	4.76
		apoptosis signaling pathway	8/102	0.191	-1.39	2.30
		oxidative stress response	4/24	0.190	-1.31	2.18
	Reactome 2016	interferon signaling	45/196	2.41e-24	-2.10	114
		eukaryotic translation elongation	32/89	7.89e-24	-2.01	107
		L13a-mediated translational silencing of ceruloplasmin expression	34/106	7.89e-24	-1.95	103
Top 1000 DETs for viral titer (874)	GO Biological Process 2015	leukocyte activation	47/373	2.50e-07	-2.38	36.2
		cellular response to cytokine stimulus	50/471	8.20e-06	-2.46	28.8
		cytokine-mediated signaling pathway	41/342	8.20e-06	-2.39	28.0
	Panther 2016	Inflammation mediated by chemokine and cytokine signaling pathway	23/188	0.000695	-1.83	13.3
		CCKR signaling map	17/165	0.0346	-1.71	5.77
		T cell activation	10/73	0.0346	-1.40	4.71
	Reactome 2016	Immune System	111/1547	0.000136	-2.23	19.9
		Interferon Signaling	26/196	0.000253	-2.09	17.4
		Cytokine Signaling in Immune system	51/620	0.00427	-2.39	13.0

Table 6.2: **Gene set enrichment analysis of DET signatures.** Abbreviations: DET, differentially expressed transcript; GO, gene ontology; CCKR, cholecystekinin receptor.

[a] Gene sets were constructed by taking the top 1,000 DETs in each category, ordered by ascending q value, and mapping to unique gene symbols

[b] q values are P-values adjusted using the Benjamini-Hochberg method.

[c] Combined scores are the product of negative log P-values and the Z-score as in Chen et al. (2013); the top three terms per annotation set, ordered by combined score, are displayed in this table.

scripts displaying divergence from the remaining distribution (Figures 6.20 and 6.21). Two of these transcripts were from *HLA-B*, one of which is its canonical protein-coding transcript, and the other of which is a retained intron; the third is the canonical transcript of *MXRA7*, which encodes a poorly characterized single-pass membrane protein. Hierarchical clustering of samples by TPM for the top 10 DETs revealed that one of the four major clusters associates with 63/80 (79%) of samples from severe cases, and the *HLA-B* transcripts strongly associate with exactly two of the three remaining clusters (Figure 6.18). Overexpression of these three transcripts in non-severe cases was consistent across both timepoints (Figure 6.23), and of the three next highly ranked transcripts, two were associated with severe cases (*CCDC144A*, *NRG1*) while one was associated with non-severe cases (*FAM69A*).

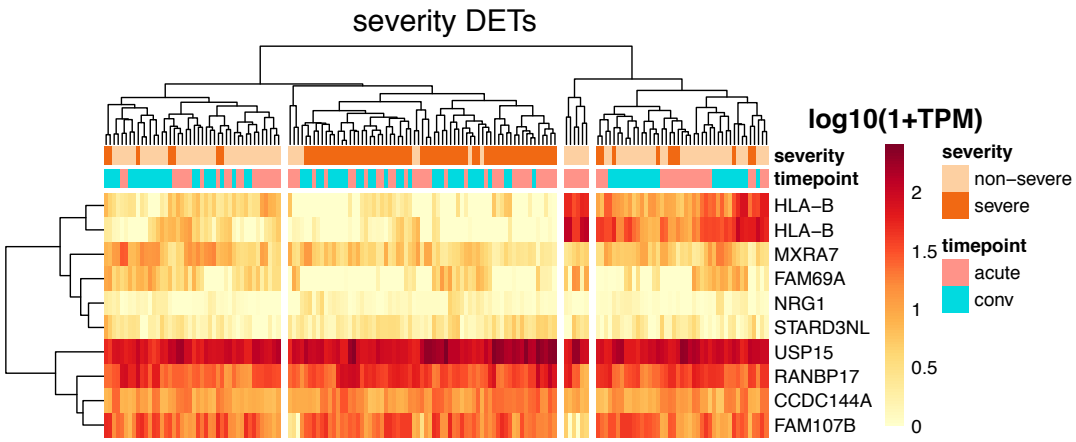


Figure 6.22: **Heatmap of DET expression** as in Figure 6.18 but for the top 10 DETs between samples from severe and non-severe cases. Four major clusters of samples are highlighted.

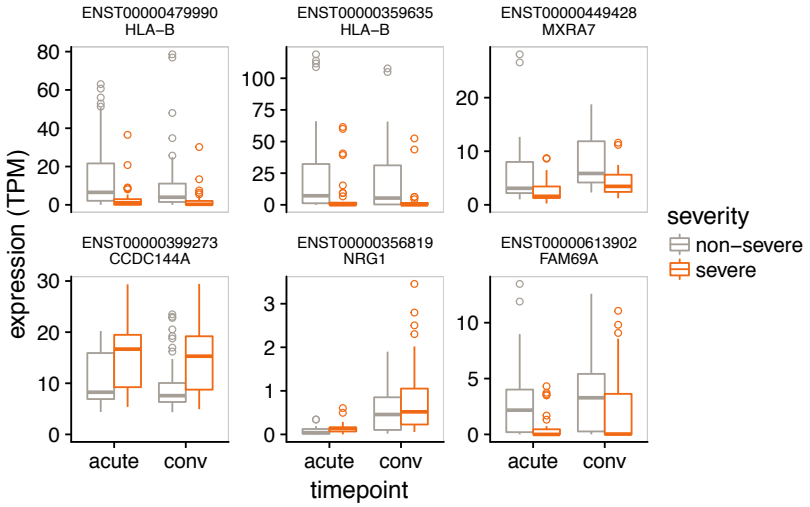


Figure 6.23: **Expression of six top DETs** (ranked by q value) between samples from severe (orange) and non-severe (gray) cases. Expression is measured in TPM. Differences in expression appear to be independent of sample timepoint (X axis).

We found a similarly sized signature for the 15d p.s.o. CHIKV IgG titer, with 63 DETs at FDR < 0.05 (Figure 6.24). Top-ranked transcripts by P value were again notable for including HLA genes, such as two transcripts of *HLA-A* and two transcripts of *HLA-DOB* among the top eight transcripts.

Complete tables of all DETs, β values, and q values for the timepoint, viral titer, symptom severity, and IgG titer contrasts are provided as Supplementary Data S3, S4, S5, and S6, respectively.

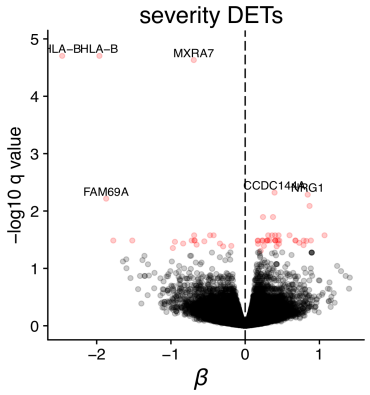


Figure 6.24: **Volcano plot** as in Figure 6.17 but for DETs between patients with higher and lower 15d post symptom onset CHIKV IgG titers. Transcripts to the right of the vertical dashed line were comparatively upregulated in patients with a higher 15d IgG, while transcripts to the left were upregulated in patients with lower 15d IgG.

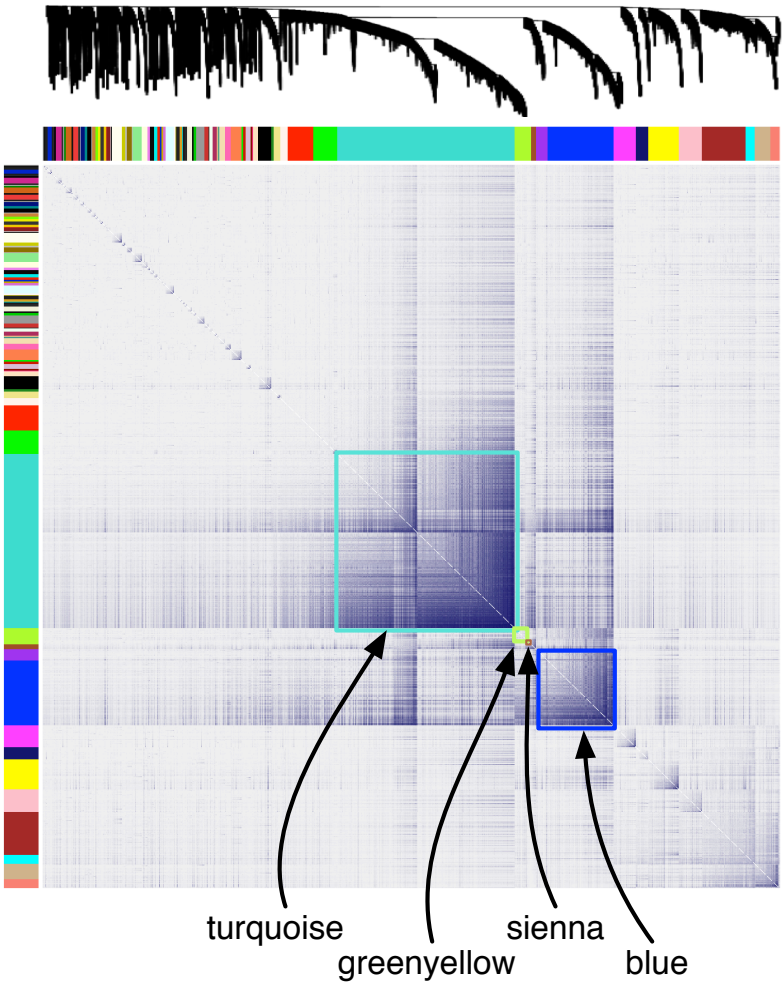


Figure 6.25: **Topological overlap matrix (TOM) plot of coexpression network** created from gene expression profiling of all 84 samples across both timepoints. At top, dendrogram of the hierarchical clustering of the matrix that undergoes a dynamic tree-cut operation to form 92 gene coxpression network modules (coEMs), depicted by the colored bars on the edges of the TOM plot (color assignment is arbitrary). Four coEMs are highlighted (boxes).

Multiscale network analysis

To relate CHIKV-associated transcriptomic changes to changes in cell subcommunity frequencies, serum cytokine concentrations, and clinical variables, we identified coexpression patterns among sets of genes to create coexpression network modules (coEMs) using whole genome coexpression network analysis (WGCNA).[44] The coEMs could then be correlated with other variables to capture the genomic coregulatory structure from biological variability present across and within the timepoints. We identified 92 coEMs, which were named after arbitrary colors (Figures 6.25 and 6.26). At a threshold of FDR < 0.05, four of these coEMs were significantly enriched for at least one of five gene sets derived from the previously acquired DET signatures for timepoint and viral

[44] Zhang and Horvath (2005), "A general framework for weighted gene co-expression network analysis."

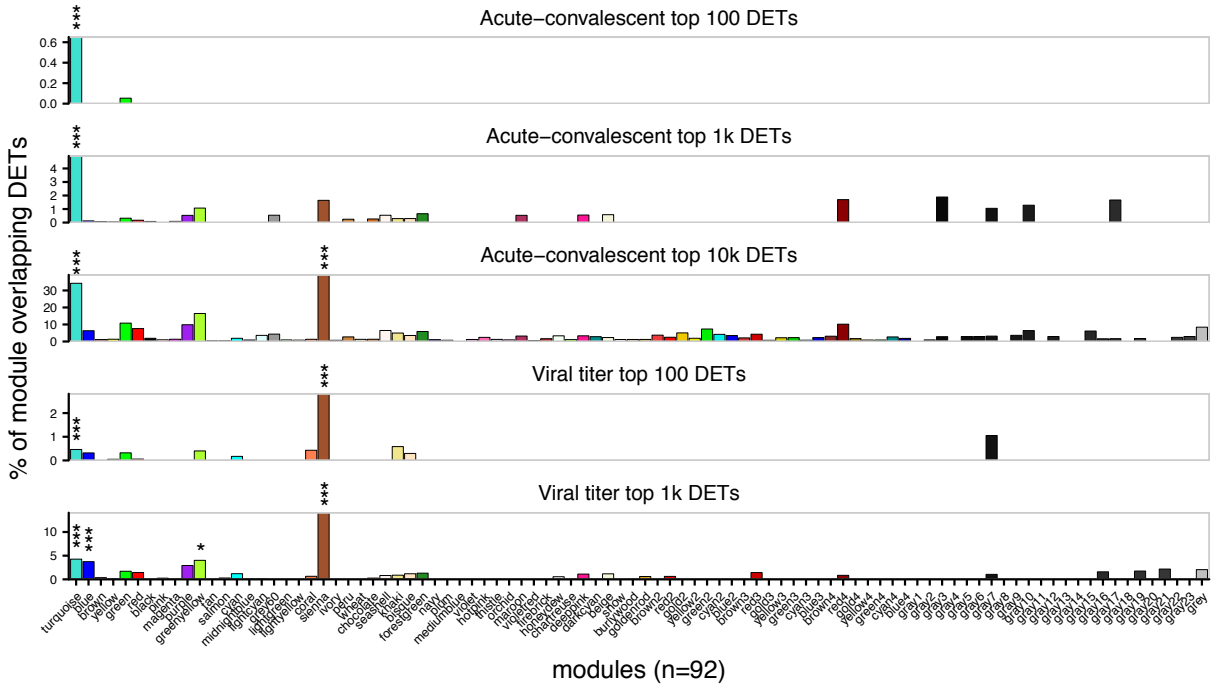


Figure 6.26: **Enrichment of five subsets of the DET signatures for CHIKV infection phase and viral titer** (see Figures 6.17 and 6.19) among each of the 92 coEMs (X axis), showing the fractional overlap of the module with the DET signature (Y axis). *q < 0.05, ***q < 0.001; q values are Benjamini-Hochberg adjusted P values (Appendix Figure B.22 shows q values for all tests). The four coEMs highlighted in Figure 6.25 have at least one significant DET signature enrichment.

titer (Figure 6.26). Of these enrichments, the most consistent and significant were turquoise (5/5 sets; max BF P = 8.0e-07) and sienna (3/5 sets; max BF P = 1.8e-08). DET enrichment q values are depicted in Appendix Figure B.22.

To explore the coregulatory structure between coEMs and clinical variables, we correlated each coEM eigengene (which is the first principal component of the expression of genes in the module) against all other coEM eigengenes and the clinical variables (Figure 6.27). This revealed that the turquoise module was strongly positively correlated with the convalescent phase (Pearson's r = 0.82), while the sienna module was strongly positively correlated with the greenyellow module (r = 0.97) and negatively correlated with the convalescent phase (r = -0.39) and turquoise module (r = -0.40). The blue module was only weakly positively correlated with the turquoise module (r = 0.37) and the convalescent phase (r = 0.26). This suggests that in our study, the sienna module is most representative of acute-associated genes, while the turquoise module is most representative of convalescent-associated genes.

We again utilized gene set enrichment analysis to explore the composition of these modules. The sienna module was most significantly enriched for GO, Re-

Gene set[a] (# genes)	Annotation set	Term	Overlap	q value[b]	Z-score	Combined score[c]
sienna (370)	GO Biological Process 2015	regulation of cytokine production	28/482	0.000268	-2.51	20.7
		positive regulation of cytokine production	22/327	0.000279	-2.45	20.1
		regulation of immune effector process	18/264	0.00106	-2.46	16.9
	Reactome 2016	immune system	65/1547	1.81e-07	-2.23	34.7
		cytokine signaling in immune system	26/620	0.0211	-2.39	9.21
		hemostasis	23/552	0.0340	-2.12	7.16
	WikiPathways 2016	type II interferon signaling	6/37	5.35e-05	-1.82	10.4
		BDNF signaling pathway	9/144	0.00141	-1.92	6.84
		senescence and autophagy in cancer	8/105	0.000720	-1.77	6.72
blue (3,554)	GO Biological Process 2015	gene expression	189/672	5.28e-08	-2.34	39.2
		hexose metabolic process	67/187	6.37e-06	-2.32	27.8
		generation of precursor metabolites and energy	107/375	0.000123	-2.37	21.3
	Reactome 2016	infectious disease	111/348	3.99e-08	-2.39	40.7
		HIV Infection	80/222	3.91e-08	-2.38	40.5
		metabolism	446/1908	3.91e-08	-2.25	38.3
	WikiPathways 2016	proteasome degradation	29/62	2.84e-05	-1.91	20.0
		B cell receptor signaling pathway	33/97	0.00455	-1.90	10.3
		pathogenic *E. coli* infection	22/55	0.00455	-1.74	9.40
greenyellow (507)	GO Biological Process 2015	negative regulation of smooth muscle cell migration	4/12	0.223	-2.65	3.97
		regulation of smooth muscle cell migration	6/31	0.223	-2.51	3.76
		endosome to lysosome transport	5/31	0.662	-2.61	1.08
	Reactome 2016	TNF signaling	5/41	0.459	-2.08	1.62
		deposition of new CENPA-containing nucleosomes at the centromere	5/52	0.459	-2.04	1.59
		nucleosome assembly	5/52	0.459	-2.03	1.58
	WikiPathways 2016	apoptosis modulation and signaling	8/93	0.353	-2.09	2.18
		complement and coagulation cascades	6/59	0.353	-1.90	1.98
		apoptosis modulation by HSP70	3/19	0.412	-1.49	1.32

Table 6.3: **Gene set enrichment analysis of coexpression modules.** Abbreviations: GO, gene ontology; BDNF, brain-derived neurotrophic factor; HIV, human immunodeficiency virus; TNF, tumor necrosis factor; CENPA, centromere protein A; HSP70, 70 kilodalton heat shock protein.

[a] Number of unique gene symbols

[b] q values are P-values adjusted using the Benjamini-Hochberg method.

[c] Combined scores are the product of negative log P-values and the Z-score as described in Chen et al. (2013); the top three terms per annotation set, ordered by combined score, are displayed here.

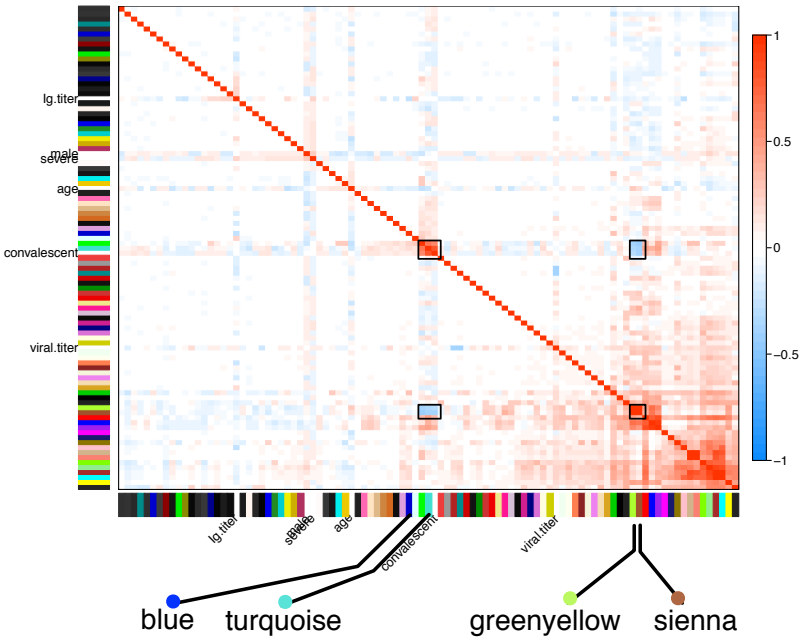


Figure 6.27: **Correlations between coEM eigengenes (colored bars on each axis) and the clinical variables (text labels).** The four coEMs from Figure 6.25 are again highlighted here. There is a strong four-way relationship between the turquoise + convalescent vs. greenyellow + sienna nodes (black boxes).

actome, and WikiPathways terms regarding the regulation of cytokine production, immune system signaling, and type II interferon signaling (Table 6.3). The blue module was significantly enriched for broader terms regarding gene expression, infectious disease, and proteasome degradation. On the other hand, the greenyellow module did not achieve any significant enrichments among these annotation libraries at FDR < 0.1, and the size of the turquoise module (10,589 genes) precluded meaningful enrichment analysis.

To create a multiscale model spanning all experimental measurements, we expanded the interaction network to include correlations with cell sub-community frequencies (from CyTOF) and serum cytokine concentrations (from Luminex). When adding the latter dataset, the large positive correlations between nearly all cytokine concentrations mentioned previously (Appendix Figures B.11–B.13) created a large connected component dominating the network structure (Appendix Figure B.23). Focusing on the transcriptomic and cell sub-community data and restricting to correlations significant at $P < 0.001$, a well-organized network formed around the primary contrast in our dataset, the acute vs. convalescent timepoints (Figure 6.28). Under a force-directed layout, cell sub-communities and gene modules that positively correlate with the convales-

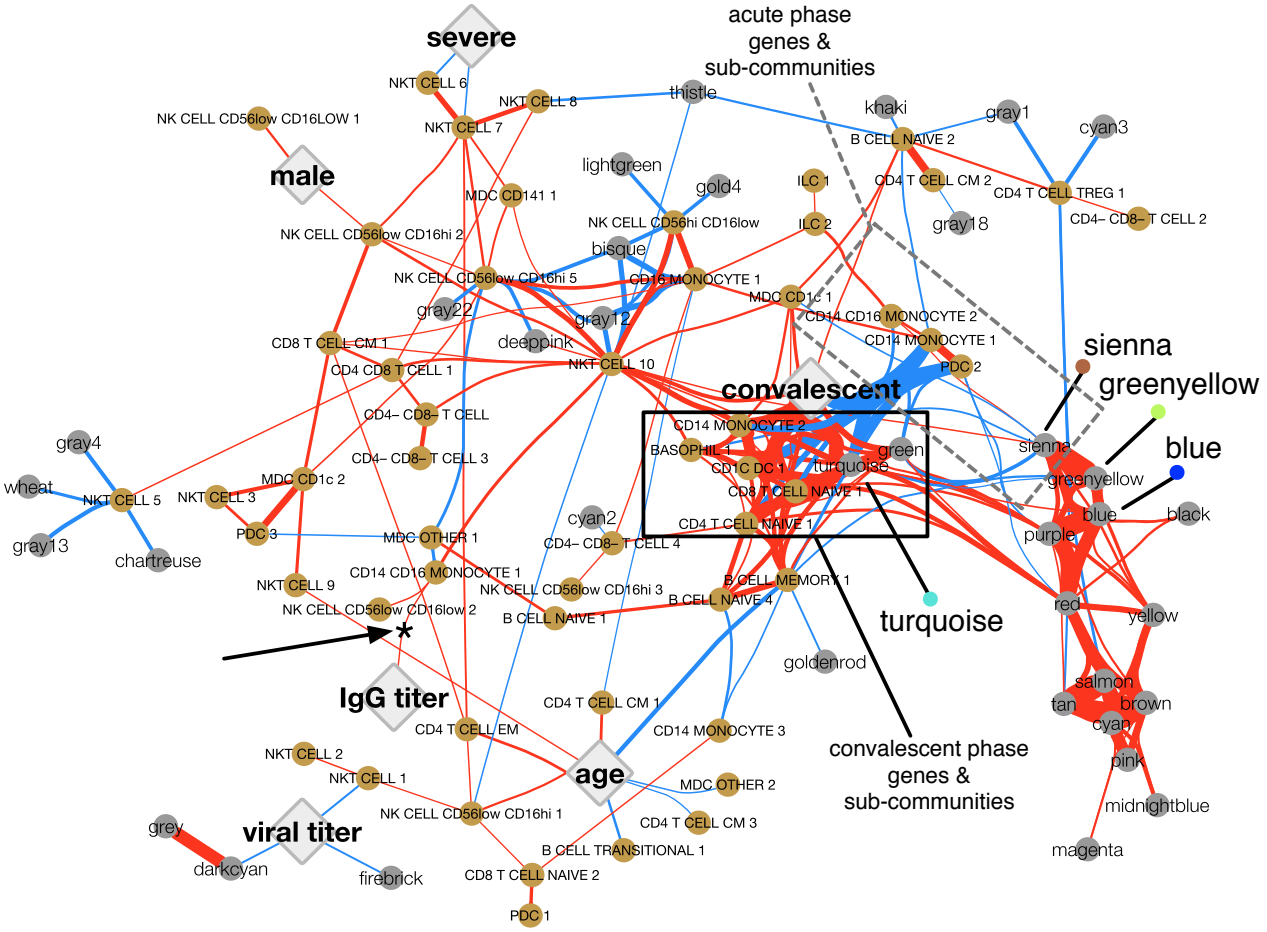


Figure 6.28: **Multiscale network of cell sub-community and coEM eigengenes depicted with a force-directed layout and edge bundling.** Gold nodes, cell sub-communities; gray nodes, coEM eigengenes; large diamonds, clinical variables; red edges, positive correlations; blue edges, negative correlations. Edges are filtered to correlations significant at $P < 0.001$, and thickness corresponds to the square of the correlations. A cluster of sub-communities and coEMs associated with convalescence (solid black box) and a corresponding cluster of sub-communities and coEMs associated with acute infection (dashed gray box) surround the "convalescent" node. A positive correlation between 15d post symptom onset CHIKV IgG titer and CD14+ CD16+ monocyte sub-community 1 (shown previously for acute phase samples in Figure 6.11) is also visible (asterisk and arrow).

cent timepoint clustered together (solid black box, Figure 6.28), while cell sub-communities and gene modules that negatively correlate with convalescence (and therefore associate with acute infection) also clustered together (dashed gray box, Figure 6.28). Weaker correlations between other gene modules, cell sub-communities, and the other clinical variables remained on the periphery of the network. For instance, the previously described significantly positive correlation between acute phase "nonclassical" $CD14^{+}CD16^{++}$ monocytes (sub-community 1) and CHIKV IgG titer reappears as a weaker positive correlation (asterisk with arrow, Figure 6.28), since the network analysis includes both timepoints.

To test whether the severe dimensionality reduction of the three datasets to the relatively small network model presented in Figure 6.28 retained predictive

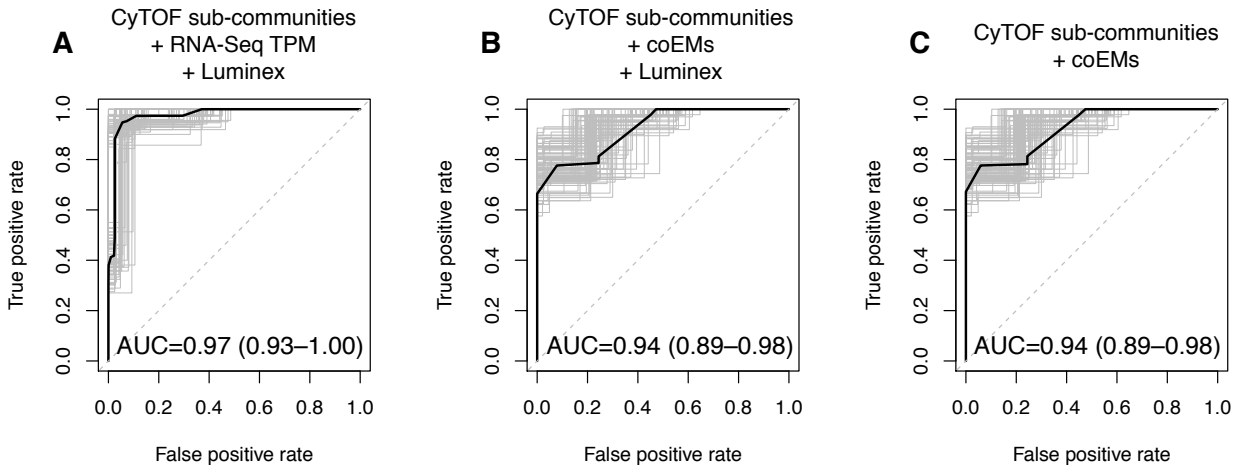


Figure 6.29: **Receiver operator characteristic (ROC) curves measuring the performance of elastic net logistic regression models predicting the phase of infection** (acute vs. convalescent) for each sample, using progressively reduced versions of the dataset. Thin grey lines show the ROC curves for 100 bootstrap replicates. The area under the curve (AUC) along with its 95% confidence interval are shown underneath each plot; a perfect classifier would achieve AUC=1 while a random classifier is expected to achieve AUC=0.5 (dashed diagonal line). A, model trained using all CyTOF sub-community frequencies, all quantified RNA-seq transcripts, and all Luminex cytokine measurements achieves near-perfect performance. B, model trained as in A with eigengene values for 92 coexpression modules replacing the RNA-seq transcript-level quantification; performance is slightly decreased but remains within the confidence interval for A's AUC. C, model trained as in B with Luminex cytokine measurements removed; performance is equivalent to the model trained in B.

value for the acute-convalescent contrast, we fit elastic net regularized logistic regression models to three versions of the merged data: A, complete per-transcript quantification, sub-community frequencies, and serum cytokine concentrations; B, same as A but with coEM eigengenes instead of per-transcript quantification; and C, same as B but with serum cytokine concentrations removed, as in the final network model. As expected, given the strong transcriptomic signature for timepoint, the model that had access to complete per-transcript quantification achieved nearly perfect performance under five-fold cross validation (area under the receiver operating characteristic [AUC]=0.97, 95% confidence interval [CI] 0.93–1.00; Figure 6.29A). Replacing per-transcript quantification with the 92 coEM eigengenes decreased predictive performance only slightly (AUC=0.94, 95% CI 0.89–0.98; Figure 6.29B). Removing the serum cytokine data, leaving only the dimensionality-reduced data used to generate Fig 7D, did not further detract from model performance (AUC=0.94, 95% CI 0.89–0.98; Figure 6.29C). This suggests that the compact, multiscale network model presented in Figure 6.28 preserves the majority of the contrast between the two phases of infection profiled by our study.

Discussion

We present in this study the most comprehensive immune profiles available for natural human infections by CHIKV. By employing a diverse CyTOF antibody panel and novel clustering techniques to systematically discover sub-commu-

nities and their frequencies in these data, we discovered more heterogeneity within PBMC populations than previously recognized in previous studies of viral infection. Considering the scarcity of any CyTOF data from viral infections,[45] these data may represent the most comprehensive immune profiles for any human viral disease. By taking an unbiased approach in measuring global changes across three scales (cell subpopulations, gene expression, and serum cytokines), we provide unprecedented robustness and detail on the effects of CHIKV on humans and a number of novel findings that compel future hypothesis-driven studies.

[45] Miner et al. (2015); Sen, Mukherjee, and Arvin (2015), "Single cell mass cytometry reveals remodeling of human T cell phenotypes by varicella zoster virus".

Strong association of CHIKV and monocyte sub-communities

On the cell population level, our findings indicate a prominent role for $CD14^+$ and $CD14^+CD16^+$ monocytes—including several novel phenotypes therein—during the acute immune response to CHIKV. Two cell sub-communities were more strongly associated with acute infection than all other sub-communities by two orders of magnitude: "intermediate" $CD14^{++}CD16^+$ monocytes and a $CD123^+$, $CX3CR1^+$ and $CD141^+$ subpopulation of $CD14^+$ monocytes.

"Intermediate" $CD14^{++}CD16^+$ monocytes are a recently described population[46] that received initial attention for showing independent predictive value for cardiovascular event risk,[47] which hinted at a role in vascular inflammation or atherosclerosis. They were also found to selectively express CCR5, the coreceptor for HIV.[48] Subsequently, studies discovered that these cells are selectively (i.e., in contrast to "nonclassical" $CD14^+CD16^{++}$ monocytes) expanded in bacterial sepsis, Dengue fever, Crohn's disease, rheumatoid arthritis, Eale's disease, and asthma.[49] Compared to our putative cluster of "nonclassical" $CD14^+CD16^{++}$ monocytes, we too saw higher expression of CCR5 in the "intermediate" subpopulation, as well as higher expression of HLA-DR, another selective marker (Appendix Figure B.3), providing strong evidence that this population is the same "intermediate" phenotype described in previous studies.[50] Very recently, in vitro studies were able to induce the "intermediate" phenotype from $CD14^+CD16^-$ monocytes by treatment with IL-10,[51] a cytokine that we also found was elevated during acute infection (Figure 6.14E). Since "intermediate" monocytes are known to be potent secretors of IL-10,[52] a positive feedback loop involving IL-10 could contribute to the strong, selective expansion in "intermediate" monocytes that was observed during the acute phase. Given how

[46] Wong et al. (2011); Ziegler-Heitbrock et al. (2010).

[47] Rogacev et al. (2012), "CD14++CD16+ monocytes independently predict cardiovascular events: A cohort study of 951 patients referred for elective coronary angiography".

[48] Ellery et al. (2007), "The CD16+ Monocyte Subset Is More Permissive to Infection and Preferentially Harbors HIV-1 In Vivo".

[49] Wong et al. (2012), "Comparable Fitness and Transmissibility between Oseltamivir-Resistant Pandemic 2009 and Seasonal H1N1 Influenza Viruses with the H275Y Neuraminidase Mutation".

[50] Zawada et al. (2011), "SuperSAGE evidence for CD14++CD16+ monocytes as a third monocyte subset."

[51] Tsukamoto et al. (2017), "CD14(bright) CD16+ intermediate monocytes are induced by interleukin-10 and positively correlate with disease activity in rheumatoid arthritis."

[52] Skrzeczyńska-Moncznik et al. (2008), "Peripheral blood CD14high CD16+ monocytes are main producers of IL-10".

much remains to be characterized about "intermediate" monocytes, and since they associate with inflammatory and autoimmune joint diseases that resemble chronic CHIKV arthropathy (like rheumatoid arthritis), this is an exciting avenue for further inquiry.[53]

Additionally, we discovered a novel subpopulation of $CD14^+$ monocytes associating with the acute phase of infection that expressed high levels of CCR4, CXCR3 and CCR6, among other markers (Figure 6.13 and Appendix Figure B.6). These markers have never been described in association with a distinct subpopulation of monocytes. This demonstrates that the heterogeneity of monocytes may extend beyond the "nonclassical", "intermediate", and "classical" divisions,[54] and that more careful identification of specific subpopulations will be needed to fully understand monocyte functionality.[55]

A globally significant correlation was found between "nonclassical" $CD14^+$ $CD16^{++}$ monocyte frequency at the acute phase and the CHIKV IgG titer two weeks later (Figure 6.11). This suggests that this subpopulation could contribute to the development of a stronger humoral response, with potentially long-term implications, given an association discovered between the early IgG response and decreased likelihood of chronic arthralgia.[56] Recent studies have suggested that monocytes can have a role in modulating activation of certain T cells,[57] which implies that monocyte subpopulations may contribute to the success of the adaptive immune response for certain pathogens. One study of this crosstalk reported induction of cytokines in δγ T cell–monocyte co-culture that closely matched the pattern of upregulation seen in our study, including IFNγ, TNFα, CCL2, IL-8, and CXCL10.[58] Even more relevant is a recent study of in vitro infection of $CD14^+$ monocytes by DENV, which upregulated CD16 expression and induced differentiation of B cells into plasmablasts, ultimately increasing IgG secretion.[59] Considering that this sequence of events closely mirrors the findings of our study, such as the high expression of CHIKV surface protein in monocytes, upregulation of $CD14^+CD16^+$ monocytes during acute infection, and the correlation with convalescent phase IgG titers, a similar mechanism may apply equally to CHIKV pathogenesis.

[53] Chaaitanya et al. (2011); Miner et al. (2015); Fingerle et al. (1993), "The novel subset of CD14+/CD16+ blood monocytes is expanded in sepsis patients."; Nakaya et al. (2012), "Gene profiling of chikungunya virus arthritis in a mouse model reveals significant overlap with rheumatoid arthritis".

[54] Appleby et al. (2013), "Sources of heterogeneity in human monocyte subsets"; Ziegler-Heitbrock and Hofer (2013), "Toward a refined definition of monocyte subsets."

[55] Stansfield and Ingram (2015), "Clinical significance of monocyte heterogeneity."

[56] Kam et al. (2012), "Early appearance of neutralizing immunoglobulin G3 antibodies is associated with chikungunya virus clearance and long-term clinical protection".

[57] Charron et al. (2015), "Monocyte : T-cell interaction regulates human T-cell activation through a CD28 / CD46 crosstalk"; Eberl et al. (2009), "A rapid crosstalk of human gammadelta T cells and monocytes drives the acute inflammation in bacterial infections".

[58] Eberl et al. (2009).

[59] Kwissa et al. (2014), "Dengue Virus Infection Induces Expansion of a CD14+ CD16+ Monocyte Population that Stimulates Plasmablast Differentiation".

Serum cytokines support a monocyte-centric response to CHIKV

In our study, most of the cytokines that were increased during acute infection were likely secreted by monocytes or contributed to the observed expansions of

monocyte subpopulations (Figure 6.14). The pronounced acute-phase increase in CXCL10, the most upregulated cytokine observed in our study, is likely linked to secretion by monocyte populations that expanded during the acute phase, since monocytes and T cells secrete CXCL10 in response to IFNγ.[60] CXCL10, which induces chemotaxis in monocytes, monocyte-derived cells, and T cells, can also be induced by IFNα and TNFα, which were upregulated in our study during the acute phase of infection.[61] Changes in CXCL10 expression are well-correlated with many infectious diseases,[62] although the role of CXCL10 in viral pathogenesis and its signalling pathways is still poorly understood (as it seems to alternately promote or protect against infection in different studies).

In the previous section, we speculated that a feed-forward loop between production of IL-10 by "intermediate" $CD14^{++}CD16^{+}$ monocytes[63] and induction of the "intermediate" phenotype by IL-10[64] could help explain the upregulation of IL-10 during acute CHIKV infection. Another contribution could be from T cells, since monocyte–T cell interactions are able to activate T cells to an IL-10 secreting state.[65] IL-10 secretion is known to be dependent on a number of regulatory factors that were transcriptionally upregulated in our study, such as *p38*, *NF-κB*, and *MyD88* (Appendix Figure B.17).[66] Another cytokine that likely contributed to the observed expansions in $CD14^{+}$ and $CD14^{+}CD16^{+}$ monocyte subpopulations is CCL2, a known $CD14^{+}$ monocyte chemoattractant,[67] which we observed to be upregulated during acute infection. CCL2 is secreted by monocytes and fibroblasts among many other cell types,[68] and like IL-10, its secretion from monocytes is also *p38* and *NF-κB* dependent.[69] CCL2 was the only cytokine that positively correlated with all monocyte subpopulation frequencies during the convalescent timepoint, which significantly differed from the correlations observed for all other cytokines (Figure 6.15). Finally, although we did not see significant upregulation of serum CCL7 levels, and the antibody panel did not include CCL8, RNA-seq did find that both of these genes were in fact transcriptionally upregulated during the acute phase (Figure 6.16). These gene products are both CCR2-binding monocyte chemoattractants (although less well characterized than CCL2) and therefore may have also contributed to expansions in monocyte populations.

[60] Luster and Ravetch (1987), "Biochemical characterization of a gamma interferon-inducible cytokine (IP-10)."

[61] Liu et al. (2011), "Combination of PB2 271A and SR Polymorphism at Positions 590/591 Is Critical for Viral Replication and Virulence of Swine Influenza Virus in Cultured Cells and In Vivo".

[62] Ibid.

[63] Skrzeczyńska-Moncznik et al. (2008).

[64] Tsukamoto et al. (2017).

[65] Charron et al. (2015).

[66] Saraiva and O'Garra (2010), "The regulation of IL-10 production by immune cells".

[67] Serbina et al. (2008), "Monocyte-mediated defense against microbial pathogens".

[68] Van Damme et al. (1994), "Induction of monocyte chemotactic proteins MCP-1 and MCP-2 in human fibroblasts and leukocytes by cytokines and cytokine inducers. Chemical synthesis of MCP-2 and development of a specific RIA.".

[69] Fietta et al. (2002), "Pharmacological Analysis of Signal Transduction Pathways Required for Mycobacterium Tuberculosis - Induced Il-8 and Mcp-1 Production in Human Peripheral Monocytes".

Most previous studies that profiled immunological changes associated

with CHIKV in humans focused on serum cytokine levels.[70] Our results were largely concordant with previous studies, but there were also some notable differences. Of the significantly elevated cytokines in our data, CXCL10, CCL2, IFNα, IL-10, and IL-1Ra concur with the changes supported by a recent systematic meta-analysis of these studies.[71] We observed two changes that were not consistent, however. An increase in IL-8 was observed during the acute phase, although previous studies vary on whether it is upregulated or downregulated in the acute phase, and it may be in fact due to a dependency of the effect on viral load.[72] We also saw an increase in TNF-α, but very few previous studies included this cytokine.

[70] Chaaitanya et al. (2011); Chow et al. (2011); Ng et al. (2009); Schilte et al. (2013), "Chikungunya Virus-associated Long-term Arthralgia: A 36-month Prospective Longitudinal Study"; Teng et al. (2015).

[71] Teng et al. (2015).

[72] Ibid.

The meta-analysis reports more cytokines that were upregulated during acute CHIKV infection than we observed, which might be expected given the increased power of combining several similarly sized previous studies. We often saw slight increases for these additional reported cytokines in our own data (e.g., IFNγ, IL-6, IL-15) but they did not reach statistical signficance in this study. One of the changes most consistently reported in the literature that was not replicated in our study was an upregulation of IL-6—we observed an upward shift that did not achieve statistical significance.[73] We also did not see globally significant associations between any cytokine levels and severity of symptoms, unlike previous studies.[74] This could be for any of a number of reasons: our statistical methods could be overly conservative, our cohort may have been too homogenous in symptomatology (the contrast between non-severe and severe cases was not emphasized during enrollment), or our cohort could simply differ too much from prior studies (adults vs. pediatric cases, differences in the CHIKV strain, or environmental differences). Reviews have already noted considerable variability in the results for serum cytokine studies on CHIKV,[75] suggesting that either much larger cohorts or a wider variety of immune profiling data will be needed to generate more reliable profiles. In part, this was a motivating factor for us to incorporate cell subpopulation and transcriptomic data into our immune profiles of CHIKV infection.

[73] Chow et al. (2011); Ng et al. (2009).

[74] Chow et al. (2011); Ng et al. (2009).

[75] Burt et al. (2017).

Transcriptomic signatures for CHIKV infection phase, viremia, severity, and immunogenicity

Our results are generally consistent with the transcriptomic signature recently reported for a C57BL/6J mouse model of CHIKV infection.[76] We both find

[76] Wilson et al. (2017).

that the strongest acute phase transcriptional upregulation occurs in interferon-associated genes like IFI's, *MX1* and *MX2*, *OAS* genes, and *RSAD2* (Viperin).[77] Notably, in mice, *CXCL10* and *CXCL9* were the most upregulated cytokine genes when comparing the acute phase to controls, with *CCL2* also substantially upregulated,[78] mirroring our measurements of the most significantly modulated serum cytokine concentrations. Although not emphasized in their study, this reveals some consistency to the monocyte-centric immune response to CHIKV between mice and humans. Our finding that type I IFN genes are not upregulated during acute infection—while initially surprising since serum concentrations of IFNα were in fact elevated—turns out to be consistent with the mouse model, which found very low RNA abundance for type I IFN transcripts.[79] Likewise, among IFN-regulated transcription factors, we find a concordant pattern of *IRF7* and *IRF9* upregulation during the acute phase, while *IRF3* is not upregulated (Appendix Figure B.21). Together, these data establish that this mouse model of CHIKV replicates many aspects of the gene expression signature induced by CHIKV in humans, including modulations of interferon pathways and monocyte-related cytokines, and therefore support its continued use as a model of human CHIKV pathogenesis and the corresponding innate immune response.

[77] Ibid.

[78] Ibid.

[79] Ibid.

Besides a large acute-convalescent transcriptomic signature, we were also able to elucidate three novel transcriptomic signatures for CHIKV viral titer, symptom severity and the convalescent phase CHIKV IgG titer. This was aided by the use of transcript-level quantification and statistical models that incorporate the uncertainty of the quantification process.[80] Notably, these signatures were constructed after incorporating timepoint as a covariate—i.e., they were found to significantly add information to a model that had already accounted for timepoint, age, and gender. Of these signatures, the strongest and least surprising was the signature for higher acute phase CHIKV viral titer, which was enriched for cytokine signalling, leukocyte activation, and interferon signaling genes. The presence of a distinct signature for viral titer in our data indicates that a more viremic acute phase must have led to transcriptional upregulation of these genes across both timepoints, otherwise they would be sufficiently captured by the acute-convalescent DET signature alone.

[80] Pimentel et al. (2016), "Differential analysis of RNA-Seq incorporating quantification uncertainty".

The transcriptional signatures for symptom severity and immunogenicity were notable for having an abundance of HLA (aka major histocompatibility

complex [MHC]) transcripts—in particular, *HLA-A*, *HLA-B*, and *HLA-DOB*. Our study showed that certain *HLA-B* transcripts appeared to be associated with less severe acute phase symptoms (Figures 6.20–6.23), while certain *HLA-A* and *HLA-DOB* transcripts were correlated with changes in the 15d CHIKV IgG titer (Figure 6.24). It is not surprising that HLA gene expression could affect infection outcomes, since the HLAs are responsible for antigen presentation and both the adaptive and cellular immune responses. For instance, during viral infection, IFNα and IFNγ increases the transcription of MHC class I loci and *HLA-B* in particular,[81] so we could speculate from our data that a particularly strong *HLA-B* response boosts the cytotoxic immune response and mitigates acute phase symptoms. Based on known roles for MHCs it is also reasonable that more transcription of *HLA-A* (a class I allele) would correlate with lower IgG titers, while more transcription of *HLA-DOB* (a class II allele) would correlate with higher IgG titers (Figure 6.24), since the former initiates the cytotoxic (non-humoral) response while the latter is involved in the adaptive and humoral responses. Our data, therefore, could be interpreted to reflect different relative prioritization of these immune responses among our cohorts' patients that manifests as a difference in magnitude of the early IgG response, which was previously observed to correlate with decreased risk of chronic arthralgia.[82] Since allelic diversity in HLA loci are well-established genetic risk factors for certain infectious and autoimmune diseases, our signatures may also reflect underlying genetic variation (e.g., expression quantitative trait loci) that affects transcription of HLA genes and thereby shifts disease outcomes.[83] Finding host genetic factors for CHIKV severity could lead to further insight into mechanisms of pathogenesis, so this is a promising direction for future study.

[81] Girdlestone (1995), "Regulation of HLA Class I Loci by Interferons".

[82] Kam et al. (2012).

[83] Kumar, Wijmenga, and Xavier (2014), "Genetics of immune-mediated disorders: From genome-wide association to molecular mechanism".

A multiscale network model of CHIKV pathogenesis

Finally, our study generated a network model that integrates global measurements of cell sub-communities, cytokines, and gene transcription into a compact roadmap of the immune responses triggered by CHIKV (Figure 6.28). A network that leverages modularity is valuable because of the inherent limitations of gene-level or cell-level analyses, which are poorly suited for traditional inference testing or Bayesian analysis because of their high dimensionality and non-independence among many of the observations. To our knowledge, we are the first to attempt a combination of WGCNA for detecting transcriptional

network modularity with comprehensive cell sub-community frequencies modeled within CyTOF data.

WGCNA produced 92 coexpression gene modules, four of which were significantly enriched for DET signatures for either infection phase or viral titer. One of these coEMs, sienna (426 genes), was significantly enriched for cytokine signalling and immune signalling terms and correlated with the acute phase of infection; a second much larger module, turquoise (10,589 genes), strongly correlated with the convalescent phase of infection. Combining gene modules with subpopulation frequencies and serum cytokine concentrations into a correlational network and filtering for edges at $P < 0.001$ produced a network dominated by intracorrelation in the cytokines (Appendix Figure B.23). Since none of the cytokines correlate significantly with the clinical variables, we removed them from the network to produce a more compact model (Figure 6.28). Under a force-directed layout, this network organizes around the primary contrast in our data—the phase of infection—with acute phase vs. convalescent phase genes and cell subpopulations separating into two communities. The sienna module also serves as a "bridge" between the timepoint contrast and most of the other strong interactions between gene modules in our dataset.

Although limited by sample size and the specific timepoints used in our study, this network represents the first completely data-driven model of the immune reaction to CHIKV across multiple layers of "omic" data. It compactly summarizes changes of hundreds of thousands of measured analytes with minimal reduction in the predictive value for the timepoint contrast (Fig 6.29), and puts these interactions into global context with other clinical variables. We hope that the generation of similar multiscale networks for other viral infections, e.g., Dengue and Zika, will soon lend insight into the comparative effects of these viruses on the human immune system and aid in the discovery of therapeutics and prognostic biomarkers that remain robust across the multiplicity of arboviral infections now prevalent in tropical urban regions.

Conclusions

Our comprehensive immune profiling of 42 pediatric cases of CHIKV infection revealed an immune response largely centered on changes in monocyte subpopulations and monocyte-related cytokines. Monocytes displayed the highest change in CHIKV surface protein expression between the two time-

points. An "intermediate" $CD14^{++}CD16^{+}$ subpopulation and an activated ($CD123^{+}$, $CX3CR1^{+}$ and $CD141^{+}$) $CD14^{+}$ monocyte subpopulation associated most strongly with the acute phase of infection when compared against all other identified subpopulations of PBMCs. Interestingly, we also found a subpopulation of $CD14^{+}$ monocytes with distinctly higher expression of previously unreported markers (CCR4, CXCR3 and CCR6) that also associated with the acute phase of infection. Although "nonclassical" $CD14^{+}CD16^{++}$ monocyte frequencies were unchanged across the timepoints, we found a significant correlation between their frequency at the acute phase and corresponding convalescent phase CHIKV IgG titers. Finally, among the elevated serum cytokine levels for the acute phase of infection, half concerned known monocyte chemoattractants (CXCL10, CCL2, and IL-10).

Our study produced additional novel findings. We confirmed that transcriptomic effects of CHIKV in humans for the different phases of infection correspond well to those recently reported for a mouse model,[84] but furthermore, we discovered new transcriptomic signatures for the level of acute phase viremia, acute phase symptom severity, and convalescent phase immunogenicity. Among the signatures for acute severity and convalescent immunogenicity, we found an abundance of specific HLA transcripts that correlated with both of these outcomes, and a notably strong correlation between severity and transcription of *MXRA7*, an essentially uncharacterized gene with only two unrelated disease associations reported in the literature.[85] We also find globally significant expression of CHIKV surface protein on several B cell subpopulations, which have never been productively infected in vitro by CHIKV.[86] Finally, we have integrated all of our observed changes into a multiscale network that summarizes the immunological changes across the cellular and gene expression levels and their interactions with certain clinical outcomes. We hope that our findings have provided a uniquely global perspective on the biomolecular landscape of CHIK pathophysiology, and that they spark new hypotheses for future experiments that can further disentangle the mechanisms of CHIKV pathogenesis and the components of a successful immune response in humans.

[84] Wilson et al. (2017).

[85] Sim et al. (2013), "Genetic Loci for Retinal Arteriolar Microcirculation"; Veiga-Castelli et al. (2010), "Genomic alterations detected by comparative genomic hybridization in ovarian endometriomas".

[86] Her et al. (2010); Sourisseau et al. (2007); Teng et al. (2012).

Materials and Methods

Study participants

To characterize immune profiles of CHIKV infection, 43 chikungunya cases (42 for analysis plus one extra) were selected from the participants aged 6 months to 14 years who were enrolled in our ongoing study at the National Pediatric Reference Hospital (HIMJR) in Managua, Nicaragua between September 2015 and April 2016. Chikungunya cases were laboratory-confirmed by detection of CHIKV using RT-PCR/virus isolation in acute-phase samples, seroconversion by IgM capture ELISA and/or a >4-fold increase in antibody titer by Inhibition ELISA in paired acute and convalescent sera.[87] Participants were also screened for dengue virus (DENV) infection, and CHIKV/DENV co-infections were excluded. To obtain a homogenous set of samples, all selected cases had an acute-phase sample collected on days 2-3 of illness and a convalescent sample collected on days 15-17 post-illness for plasma, whole blood in PAXgene solution, and peripheral blood mononuclear cells (PBMCs). PBMCs were isolated from whole blood as previously described.[88] Clinical information was collected every 12 hours, and was digitized by double-data entry under systematic monitoring by a clinical supervisor, with quality control checks performed daily and weekly.

[87] Galo et al. (2017), "Development of in-house serological methods for diagnosis and surveillance of chikungunya".

[88] Zompi et al. (2012), "Dominant cross-reactive B cell response during secondary acute dengue virus infection in humans."

CyTOF sample processing and acquisition

Cryopreserved PBMC samples from the acute and convalescent phases of infection were thawed and stained with Rh103 nucleic acid intercalator (Fluidigm) as a viability marker. Paired PBMC samples from each timepoint were first barcoded using a CD45 antibody-based barcoding approach,[89] and each acute and convalescent sample pair was pooled as a single patient sample for subsequent processing to minimize technical variability and potential batch effects. The pooled patient samples were then stained with a validated 37-marker CyTOF antibody panel (Appendix Table A.2) for 30 minutes on ice and then fixed, permeabilized and incubated with Ir nucleic acid intercalator (Fluidigm). The samples were then stored in freshly-diluted 2% formaldehyde in PBS and stored until acquisition. Immediately prior to CyTOF acquisition, the samples were washed with deionized water (diH20), counted and resuspended in diH20 con-

[89] Mei, Leipold, and Maecker (2016), "Platinum-conjugated antibodies for application in mass cytometry".

taining a 1/20 dilution of Eq 4 Element beads (Fluidigm). Following routine autotuning, the samples were acquired on a CyTOF2 mass cytometer (Fluidigm) equipped with a SuperSampler fluidics system (Victorian Airships) at an event rate of <400 Hz.

CyTOF data analysis

Following data acquisition, the FCS files were normalized using the bead-based normalization algorithm in the CyTOF control software, and uploaded to Cytobank for initial data processing. Normalization beads were excluded based on Ce140 signal, and cell events were identified based on Ir191/193 DNA signal. A conservative doublet exclusion gate was applied based on DNA and event length, and $Rh103^{+}$ dead cells were also excluded. The cell events associated with the acute and convalescent samples were then manually de-barcoded based on CD45-194Pt and CD45-198Pt expression, respectively, and were split and exported as separate samples for subsequent analyses using a semi-supervised computational analysis pipeline.

We first applied traditional hierarchical gating to a subset of samples to identify 9 major immune compartments: $CD4^{+}$ and $CD8^{+}$ T cells, B cells, NK cells, NKT cells, Monocytes, mDCs, pDCs, and basophils. This manually gated data was used to train a logistic regression classifier (which we term `Nod`), which was then applied to identify these populations in all the samples. We then applied Phenograph[90] as an unbiased approach to define the phenotypic heterogeneity within each of these compartments (`HybridLouvain`). The cell clusters identified in each single sample were then meta-clustered across all samples to identify phenotypically-similar communities that were reproducibly present across multiple samples (`MetaHybridLouvain`). These meta-clusters were then manually annotated based on overall marker expression profiles and their association with known immune cell subsets, allowing for the presence of additional phenotypically distinct sub-clusters within these known subsets. These annotations were mapped back to the individual samples, and the relative frequency and median marker expression patterns of these consistently annotated clusters were then exported for further statistical analyses. Metaclusters that were characterized by protein expression patterns that did not correspond to any known cell subsets, including those that appeared to be cell-cell doublets, were annotated as "undefined" and not included in subsequent statistical or network analyses.

[90] Levine et al. (2015).

Multiplex ELISA

Cytokines and chemokines were measured using a multiplex ELISA-based assay (Luminex). All serum samples were inactivated with a UV-C lamp (254 nm) for 10 min on ice in a biosafety-level 3 laboratory at the University of California, Berkeley. Each sample was run in duplicate in a 96-well micro titer plate using 25 µL of plasma or serum from each patient from acute and convalescent time points using the multiplex cytokine panels (Multiplex High Sensitivity Human Cytokine Panel, Millipore Corp.). 41 analytes (cytokines and chemokines) were measured measured using a Luminex-200 system and the XMap Platform (Luminex Corporation). Acquired mean fluorescence data was analyzed and calculated by the Beadview software. The lower and upper detection limits for these assays are 3.0 pg/mL and 15,000 pg/mL, respectively. Quality control of each sample was performed and samples with bead counts <50 were not used for analysis.

Viral titer assays

Viral RNA was extracted from 140 µL of cell culture supernatant or 140 µL of patient serum using the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's protocol, and RNA was eluted in 60 µL of RNase-free water. Primers for the E1 gene were designed to quantify CHIKV copies in each patient and were used at 300 nM final concentration. The forward primer is 5'-CATCTGCACYCAAGTGTACCA-3' and the reverse primer is 5'-GCGCATTTTGCCTTCGTAATG-3'. A TaqMan labeled probe was used for detection: FAM-5'-GCGGTGTACACTGCCTGTGACYGC-3'-BHQ-1.[91] The SuperScript III One-Step RT-PCR System (Invitrogen) was used for reverse transcription of viral RNA and subsequent amplification of viral complementary DNA (cDNA). Specifically, 5 µL of extracted viral RNA, 0.5 µL of SuperScript III RT/Platinum Taq High Fidelity Enzyme Mix (Invitrogen), 12.5 µL of 2× Reaction buffer, 5 µL RNase free water and 2 µL of primers and probes were added to each well. Viral RNA was reverse transcribed (52°C for 15 minutes), and the resulting cDNA was amplified via one cycle of denaturation (94°C for 2 minutes), 45 cycles of denaturation (94°C for 15 seconds), annealing (55°C for 40 seconds), and extension (68°C for 10 seconds). For quantitation of CHIKV copies, a 4-point standard curve (8.0, 6.0, 4.0, and 2.0 $\log_{10}$ copies/µL of eluate) was used. Standard curves were prepared

[91] Waggoner et al. (2016), "Viremia and Clinical Presentation in Nicaraguan Patients Infected with Zika Virus, Chikungunya Virus, and Dengue Virus."

using quantitated ssDNA (Integrated DNA Technologies) containing the target sequence of CHIKV with the following sequence: 5'-CACAACATCTGCACCCAAGTG TACCACAAAAGTATCTCCAGGCGGTGTACACTGCCTGTGACCGCCATTGTGTCATCGTTGCATT ACGAAGGCAAAATGCGCACTAC-3'

Preparation of RNA sequencing libraries

Total RNA was extracted from PaxGene RNA blood with the PAXgene Blood RNA Kit (Qiagen) by following manufacturers' instructions including DNAse digestion and an additional clean-up using RNEasy MinElute kit (Qiagen). Purified RNA samples were quantified by Qubit 3.0 fluorometer with RNA High Sensitivity Assay kit (Thermo-Fisher). The quality of the RNA was confirmed by TapeStation 2200 with the RNA High Sensitivity ScreenTape (Agilent Technologies). To prepare the 84 samples' libraries, ribosomal RNA (rRNA) and globin mRNA were removed from 200ng total RNA, and the remaining RNA was fragmented and primed for cDNA synthesis using TruSeq Total Stranded RNA HT kit with Ribo-Zero Globin on a Microlab STAR automated liquid handling system (Hamilton). The libraries were barcoded with TruSeq HT indices to allow for multiplexing and ligation-mediated PCR was performed to enrich barcoded libraries for 15 cycles, then purified with the Agencourt AMPure XP beads system (Beckman Coulter). The libraries were assessed for quality with the high sensitivity DNA chip in a TapeStation 2200 (Agilent) and quantified with KAPA Library Quantification Kits for Illumina platforms (Kapa Biosystems). The libraries were diluted to 2nM and combined equimolarly in pools of 12. These pools were then clustered using a cBot (Illumina) with a HiSeq 3000/4000 paired-end cluster kit on a patterned flow cell, one pool per lane. The flow cell was sequenced on a HiSeq4000 using a HiSeq 3000/4000 SBS kit (300 cycles, Illumina). Two technical replicates were sequenced per biological sample for a total of 168 sequencing runs.

Pre-processing of RNA-seq data

Sequencer-generated base call (BCL) files were converted to FASTQ files and the multiplexed samples were separated using `bcl2fastq`, which was then assessed for sequencing quality using FastQC[92] (version 0.11.4) to assess the quality of the sequencing data. The FASTQ files were quality filtered by using FASTAX Toolkit[93] with the invocation `fastq_quality_filter -q 30 -p 50 -v -Q`

[92] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[93] http://hannonlab.cshl.edu/fastx_toolkit/

33, and only the sequencing reads that met all quality control requirements were aligned to the latest human reference genome (GRCh38) using HISAT2[94] (version 2.0.4). SAMtools[95] (version 0.1.19) was used to sort and convert the SAM files to BAM. Aligned sequences were assembled into potential transcripts using StringTie[96] (version 1.2.2). SAM files of preprocessed RNA-seq alignments were also analyzed with the `htseq-count` script from HTseq[97] for differential expression analysis by counting the overlap of reads with genes.

[94] Kim, Langmead, and Salzberg (2015), "HISAT: a fast spliced aligner with low memory requirements."

[95] Li et al. (2009), "The Sequence Alignment/Map format and SAMtools".

[96] Pertea et al. (2015), "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads."

[97] Anders, Pyl, and Huber (2015), "HTSeq-A Python framework to work with high-throughput sequencing data".

Differential expression analyses

The assembled transcripts, representing the different splice variants for each gene locus, were quantified in order to analyze differential expression in response to CHIKV infection. Differentially expressed genes were analyzed using the R packages `ballgown`[98] and `DESeq2`,[99] and the results were consolidated to increased robustness. The samples were compared between acute and convalescent timepoints with gender, age, and infection severity included as covariates. Pathway-based visualization of differentially expressed genes was performed with the `pathview`[100] R package and KEGG[101] annotations.

For differential expression analysis at the transcript level, we used `kallisto`[102] (version 0.43.0) and `sleuth`[103] (version 0.28.1). These methods, based on pseudoalignment counts, are more computationally efficient than complete alignment and can use bootstrap replicates to model uncertainty at the quantification step.[104] Uncertainty in quantification is caused by the inherent ambiguity of assigning reads to transcripts and constrains other methods that use invariant transcript counts, whereas `sleuth`'s method can maintain sensitivity at the isoform level while adequately controlling the FDR.[105] Pseudoalignment utilized a transcriptome index built from Ensembl release 79 (March 2015) for GRCh38. `sleuth` uses an additive response error model under which each variable or covariate has fixed effects β on each transcript's abundance, and to determine significance, the Wald test was used for the null hypothesis that $\beta_t = 0$ for each transcript t. For the acute vs. convalescent (timepoint) signature, age and gender were included as covariates in the model, while for the viral titer, severity, and 15d IgG titer signatures, timepoint, age, and gender were included as covariates in the model. Viral titers and 15d IgG titers were modeled in units of $\log_{10}$ dilutions. Inference tests for differential expression were adjusted for multiple hypotheses using the Benjamini-Hochberg procedure.[106]

[98] Frazee et al. (2014), "Flexible analysis of transcriptome assemblies with Ballgown"; Pertea et al. (2016), "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown".

[99] Anders and Huber (2010), "Differential expression analysis for sequence count data."

[100] Luo and Brouwer (2013), "Pathview: An R/Bioconductor package for pathway-based data integration and visualization".

[101] Ogata et al. (1999).

[102] Bray et al. (2016), "Near-optimal probabilistic RNA-seq quantification".

[103] Pimentel et al. (2016).

[104] Bray et al. (2016).

[105] Pimentel et al. (2016).

[106] Benjamini and Yekutieli (2001), "The control of the false discovery rate in multiple testing under depencency".

Construction of gene coexpression networks and coexpression modules

Gene coexpression networks were constructed from the gene-level expression data for all samples using weighted gene coexpression network analysis (WGCNA) using the `WGCNA`[107] (version 1.51) and `coexpp`[108] (version 0.1.0) R packages. WGCNA leverages natural variance in expression between sampled individuals and timepoints to build a network structure from the Pearson correlations for all gene-gene pairs.[109] `coexpp` is a specialized implementation of WGCNA that optimizes memory and multicore usage. Gene expression data were preprocessed for WGCNA by applying a $\log_2$ transformation to the FPKM quantification and removing the lowest-variance quartile of genes.[110] In our study, we did not normalize each timepoint separately, as we were interested in correlations across the timepoints and interactions between gene modules and the phase of infection.[111] WGCNA then converts the gene-gene correlation matrix into an adjacency matrix using a power function that optimizes for scale-free topology, and adjacencies are then transformed into a topological overlap matrix (TOM) that represents normalized counts of neighbors that are shared between the nodes on either side of each edge. Genes were grouped using average-linkage hierarchical clustering of the TOM, followed by a dynamic cut-tree algorithm[112] that divides the dendrogram branches into gene coexpression network modules (coEMs). Relationships among coEMs and the other data were evaluated using eigengenes[113] (the first principal component of each coEM), calculating the Pearson correlations for all possible pairings of the coEM eigengenes, clinical variables, and cell subpopulations. Network layout was performed using the ForceAtlas2 algorithm in Gephi[114] (version 0.9.1) followed by visualization in Cytoscape[115] (version 3.4.0).

[107] Zhang and Horvath (2005).

[108] `https://bitbucket.org/multiscale/coexpp`

[109] Zhang and Horvath (2005).

[110] WGCNA's use of Pearson's r does not require assuming the data are normally distributed. Nefzger and Drasgow (1957) note that this is a needless assumption, and Edgell and Noon (1984) demonstrate that Pearson's r is generally robust to non-normality for $N \geq 15$. Furthermore, Allen et al. (2012) show that with non-normal (bimodal) data, WGCNA still outperforms Bayesian and partial correlation network reconstruction methods.

[111] A network describing correlations within each timepoint that remain unchanged across timepoints would result from standardizing acute and convalescent data separately. For more details on preparing RNA-seq and heterogeneous data for WGCNA, see Langfelder and Horvath's FAQ at `https://goo.gl/LU88Df`.

[112] Langfelder, Zhang, and Horvath (2008), "Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R".

[113] Langfelder and Horvath (2007), "Eigengene networks for studying the relationships between co-expression modules".

[114] Bastian, Heymann, and Jacomy (2009), "Gephi: An Open Source Software for Exploring and Manipulating Networks".

[115] Smoot et al. (2011), "Cytoscape 2.8: new features for data integration and network visualization."

Gene set enrichment analyses

The acute-convalescent and viral titer DET signatures were analyzed for enrichment of Gene Ontology (GO) biological process[116] (2015), Panther[117] (2016), and Reactome[118] (2016) terms using the Enrichr platform.[119] DETs were ranked by q value, and query sets for Enrichr were created from the 100, 300, 1,000, and 3,000 top ranked DETs (all having $q < 0.05$) mapped to unique gene symbols, which all produced qualitatively similar results for top enriched

[116] The Gene Ontology Consortium (2015), "Gene ontology consortium: Going forward"

[117] Mi, Muruganujan, and Thomas (2013), "PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees".

[118] Fabregat et al. (2016), "The reactome pathway knowledgebase".

[119] Chen et al. (2013), "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool."

terms; representative results for 1,000 DETs are presented in this study. Enrichr improves on the typical method of ranking term significance with one-sided Fisher's exact tests by multiplying their log-scaled P values by a Z-score of the deviation from the expected rank for each term, which decreases the bias of the Fisher's exact method toward terms with few gene assignments.[120] Enrichment of WGCNA coEMs for terms from GO biological process (2015), Reactome (2016), and WikiPathways[121] (2016) was similarly calculated using Enrichr without ranking or cutoffs. Enrichment of DET signatures within each coEM was calculated using one-sided Fisher's exact tests and a Benjamini-Hochberg adjustment.

[120] Ibid.

[121] Kutmon et al. (2016), "WikiPathways: Capturing the full diversity of pathway knowledge".

Data availability

Raw RNA-seq reads and differentially expressed genes were deposited in GEO under study accession number GSE99992.

Statistical analyses

Inference testing for unpaired quantitative variables[122] (CyTOF community and sub-community frequencies, marker expression means per sub-community per sample, and correlations between monocyte sub-community frequencies and Luminex analyte concentrations) used the Mann-Whitney U, while the Wilcoxon signed-rank test was used for paired comparisons (Luminex analyte concentrations). Hypothesis testing for correlations used either Spearman's ρ (sub-community frequencies vs. viral and IgG titers) or Pearson's r (multiscale network analysis). Visualization of small correlation matrices was performed with the `corrplot` R package. P values in this study were adjusted for multiple hypotheses using either the Bonferroni or Benjamini-Hochberg (FDR-based aka q value) methods, as specified throughout the Results. Elastic net regularized regression, which fits a logistic regression model with ℓ_1 and ℓ_1 penalties (the elastic net penalty), was performed with the `glmnet`[123] R package (version 2.0-5). Elastic net hyper-parameters α and λ were both selected empirically per model by a grid search that maximized AUC under five-fold nested cross validation. 100 bootstrap resampling runs were used to estimate the 95% confidence interval for the AUC. R version 3.2.2 was used for all statistical analyses, and in addition to those already mentioned, the following package versions were used: `ggplot2` (2.2.1), `pheatmap` (1.0.8), `ROCR`[124] (1.0-7), and `Biobase` (2.30.0).

[122] Although the study design is paired across timepoints, `MetaHybridLouvain`'s inability to identify every sub-community in all samples creates null values that complicate traditional paired analysis. Additionally, it is not possible to regress traditional linear models if patient ID is a covariate while also modeling the effects of per-patient covariates like gender, severity, and IgG titer (the design matrix will be rank deficient). One potential solution is to used mixed-effect models as implemented in limma-voom—see Ritchie et al. (2015) and Law et al. (2014)—which can model a random effect (as a varying intercept) for differences between patients. The downsides of this approach, both of which can result in decreased power, are the additional model degrees of freedom and limma-voom's inability to model the uncertainty of the quantification step provided by the `kallisto` bootstraps. Nevertheless, a repeat analysis using these models as implemented in the limma-voom framework by collaborators Mayte Suarez-Fariñas and Maria Suprun has recapitulated the major results of this chapter, and will be included in the submitted publication.

[123] Friedman, Hastie, and Tibshirani (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent".

[124] Sing et al. (2005), "ROCR: Visualizing classifier performance in R".

Notes

Contributors

Theodore R. Pak (TRP), Daniela Michlmayr (DM), Adeeb H. Rahman (AHR), El-Ad David Amir (EDA), Eun-Young Kim (EK), Angel Balmaseda (AB), Seunghee Kim-Schulze (SKS), Michael G. Stewart (MGS), Guajira P. Thomas (GPT), Steven Wolinsky (SW), Andrew Kasarskis (AK), and Eva Harris (EH) contributed to this chapter.

TRP, DM, AHR, SW, AK, and EH designed the study. EH and AB directed studies in Nicaragua to obtain the samples for this study. AHR and SKS performed the CyTOF experiments. AHR, SKS, and EDA performed manual gating of CyTOF data, clustered events with `MetaHybridLouvain`, and created the viSNE plots within Figures 6.2, 6.6, and 6.8. AHR wrote the first draft of the CyTOF Methods subsections. EH and DM selected study participants; DM analyzed demographic data, performed the viral titer assays, and wrote the first draft of the corresponding Methods subsections and Table 6.8. DM and SKS performed Luminex assays. EK, MGS, and GPT prepared RNA-seq libraries, performed the sequencing, pre-processed the data, analyzed differentially expressed genes, and wrote the first draft of the corresponding Methods subsections. TRP analyzed differentially expressed transcripts, constructed network and predictive models, and performed statistical analyses. TRP created all other figures in this chapter and the Appendix and wrote the first draft of all other sections of this chapter. TRP and DM share first authorship on the manuscript for submission.

Funding

This work was supported by NIH/NIAID grants U19-AI118610 (TRP, DM, EH, AK, AHR, SKS, SW, EK, AB), R33-AI100186 (AB, EH), F30-AI122673 (TRP), and in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Jesse Waggoner for advice regarding quantifying CHIKV viremia in patient samples. We thank study personnel at the HIMJR and the National Virology Laboratory of the Ministry of Health in Managua, Nicaragua for enrolling patients, collecting blood samples and clinical data, maintaining databases and a high level of quality con-

trol, and preparing PBMCs. We are grateful to the study participants and their families. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

Conflict of Interest

The authors have no conflicts of interest to disclose.

Acknowledgements

This work was supported in part by the resources and expertise of the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Mayte Suarez-Fariñas and Maria Suprun for performing a confirmatory analysis using alternative statistical methods (such as mixed-effect models) and incorporating data from a repeat Luminex experiment, which reaffirmed the major findings in this chapter after the dissertation was defended. We thank study personnel at the HIMJR and the National Virology Laboratory of the Ministry of Health in Managua, Nicaragua for enrolling patients, collecting blood samples and clinical data, maintaining databases and a high level of quality control, and preparing PBMCs. We are grateful to the study participants and their families. We thank Ana Fernandez-Sesma and Irene Ramos-Lopez and all members of the Dengue Human Immune Profiling Consortium for their assistance in designing and conducting the study and their feedback on preliminary results.

7

Discussion and Conclusions

What I do, Bill, is I catch the world in the headlights of my justice.

—STEPHEN COLBERT

Um, I'll tell you the problem with the scientific power that you're using here. It didn't require any discipline to attain it. You read what others had done and you took the next step. You didn't earn the knowledge for yourselves, so you don't take any responsibility for it. You stood on the shoulders of geniuses to accomplish something as fast as you could, and before you even knew what you had, you patented it, and packaged it, and slapped it on a plastic lunchbox, and now ... [bangs on the table]

—IAN MALCOLM, *Jurassic Park*

THIS DISSERTATION addressed several frontiers of multiscale analysis—in this context, the integration of varied sources of omic and clinical informatics data—wherein we sought new types of actionable knowledge that infectious disease clinicians could find useful. In this final chapter, I review some of the major findings from these chapters, reflect on lessons learned from these projects in total, and finally consider the most promising directions for future work.

Summary of major findings

Chapter 1 introduced a futuristic vision for bringing multiscale analysis into clinical infectious diseases, which included the routine use of next-generation sequencing (NGS) on hospital pathogen isolates and the construction of databases and software to integratively analyze these data in the context of electronic medical records (the EMR). A relatively small-scale application of these principles was presented in Chapter 2, which was a study that admittedly resulted from ad-hoc rather than routine analysis, but which was nevertheless well suited for piloting Mount Sinai's Pathogen Surveillance Program (PSP). From the PSP's perspective, it importantly demonstrated that our team had the capability to rapidly generate complete genomes for hospital isolates, which led to the initiation of routine collection and sequencing of isolates for more common hospital-associated infections (HAIs). Although the original intent of the pilot study was to search for relatedness between the sequenced *Stenotrophomonas maltophilia* isolates, as they were collected as a result of a perceived "spike" in

hospital-wide incidence, the NGS data quickly established that none of them were related: 1,000s of single nucleotide variants (SNVs) separated all different-patient isolates.

Rather fortuitously, multiple serial isolates had been collected from one of the patients, and they surrounded a change in antimicrobial susceptibilities measured by the Clinical Microbiology Lab using routine automated culturing techniques (Vitek). Because we had already sequenced and assembled the isolates, we had a perfect opportunity to find genetic changes that corresponded to those phenotypic shifts. As Chapter 2 revealed, I was in fact able to trace the emergence of quinolone resistance to a SNV in the *smeT* gene, which is a known *tetR*-like repressor for a membrane pump that effluxes quinolones. Since meticulous in vitro studies of sequence changes in this gene under selective pressure were fortuitously available in the literature,[1] I was able to match this SNV with an identical, previously characterized variant that was proven to confer quinolone resistance. In summary, the study ultimately showed that enumerating the mechanisms of emerging drug resistance in hospitalized bacteremias could be a suitable secondary goal for PSP analyses, beyond simply surveilling for patient-to-patient transmissions.

[1] Alonso and Martínez (1997), "Multiple antibiotic resistance in Stenotrophomonas maltophilia."; Sánchez, Alonso, and Martinez (2002), "Cloning and characterization of SmeT, a repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF".

Chapters 3 and 4 illustrate the details of software designed by the author which now serve core steps of the PSP workflow with the contributions of additional code by other PSP members and colleagues at the Roth laboratory.[2] One of the challenges of creating so many new bacterial genomes as would be expected for any PSP that chooses long-read sequencing and *de novo* assembly is the effective vetting and collaborative analysis of these data, as there are essentially no online tools for this purpose. In Chapter 3, I showed that an web-based genome browser (ChromoZoom) can in fact dynamically load all the relevant data formats for a custom genome assembly and draw their contents in real-time within a web browser using HTML5 technologies. No other current browsers are able to load custom genome assemblies and rich annotations in this manner. In revamping ChromoZoom to do this, I ensured that it can still display curated annotations from UCSC for standard vertebrate assemblies alongside user-provided alignment and variant tracks, as this remains a common use case for other genome browsers. Browsing vertebrate genomes is also important for general microbiology research when visualizing host gene expression and/or variant data that relates to different immune responses, as

[2] At the University of Toronto and Mount Sinai Hospital in Toronto, ON; no relation to The Mount Sinai Hospital in New York, NY. See `http://llama.mshri.on.ca/`

introduced in Chapter 6.

In Chapter 4, I presented a new software suite called PathogenDB that implements the core vision of Chapter 1, which was the design of a learning healthcare system for infectious diseases depicted in a circular diagram (Figure 1.2). In creating this suite, we essentially decomposed Figure 1.2 into separate modules for NGS assembly and annotation, the comparative analysis of the resulting genomes, and delivering visualizations/reports to clinicians and implemented these modules as the software packages PathogenDB-pipeline, PathogenDB-comparison, and PathogenDB-viz. We demonstrated that diligent engineering can produce a pipeline that outputs hundreds of completed and annotated genomes, usually within a few hours of the delivery of the raw sequencing reads. As forecasted in Chapter 1, this was less a matter of inventing new tools and more a challenge of encapsulating existing tools into discrete chainable steps that track their dependencies and gracefully handle errors, and sometimes, swapping in more mature tools as they are released.[3] Indeed, the current bottleneck for the PSP is the throughput of Mount Sinai's Genomics Core Facility (which only has so many PacBio RS II machines and many other active projects) and not bioinformatics analyses. This is a massive advance over the early days of the PSP, where all steps for analyzing sequencing data were performed manually and individual investigations took months.

[3] Seemann (2014), "Prokka: Rapid prokaryotic genome annotation"; Hunt et al. (2015), "Circlator: automated circularization of genome assemblies using long sequencing reads".

The use of the PathogenDB suite now enables active monitoring of *C. difficile* and *S. aureus* isolates collected from all culture or toxin positive specimens at Mount Sinai. Despite the enormous amounts of data generated by sequencing all of these isolates, I can tidily summarize the current status of transmissions supported by NGS evidence using PathogenDB-viz. Presented as the final figures in Chapter 4, I reduce the genetic distances between hundreds of isolates into a filterable, organized heatmap. From the viewpoint of an interpreting clinician, the simple presence and size of brightly colored squares in this diagram is the hallmark of a past or present outbreak. Should we be able to reduce the turnaround time of samples through our Genomics Core Facility (currently, several weeks), we would be able to monitor and track outbreaks of HAIs in real time with near exactitude, given the level of certainty provided by NGS. Even in spite of the semi-realtime nature of the current process, it has been used to confirm one large outbreak of *S. aureus* that was previously suspected by infection control personnel, and also discovered a new (genetically

unrelated) 8-patient outbreak that infection control did not know of, as it was spread over many units and several months. Although the investigation into the cause of the second outbreak is still ongoing, one notable lead is that the index patient appears to have visited Mount Sinai many times throughout the time of the outbreak, transiting through several units on each visit. Analyses of severe acute respiratory syndrome (SARS) re-surfaced the concept of "super-spreaders," i.e., infected individuals that transmit to disproportionate numbers of secondary contacts,[4] and this pattern has already been observed in at least one NGS study of hospital-acquired *S. aureus*.[5] This may be the role of the index patient in our new cluster, and based on the PSP's analysis, infection control personnel at Mount Sinai have already initiated special monitoring and isolation precautions for this patient to prevent further transmissions.

[4] Stein (2011), "Super-spreaders in infectious diseases".

[5] Tong et al. (2015), "Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting".

This introduces the question of how to rationally select interventions based on any proof of transmissions by NGS data, even if we eventually gain the capacity to do this for hospital pathogens in real time. Chapter 5 attempts to address a foundational aspect of this problem by estimating the local cost of one HAI, *C. difficile* infection (CDI), using only data that is readily available in the EMR. Without knowing this local cost, it is impossible to perform a cost-benefit analysis for any proposed intervention, ranging from relatively inexpensive bed cleanings to extremely expensive hypothetical programs that sample and sequence the entire environment of the hospital to find reservoirs. While it is possible to simply rely on national estimates of per-patient cost for common HAIs, these estimates have proven to be highly variable, sometimes by two orders of magnitude. Also, costs often cannot be estimated directly from billing data due to historical divides between these data and EMRs in most hospitals and the wayward incentives (in the US system) that cause billed prices to completely lack any correlation with operational costs.[6] Based on rigorous, reproducible analysis of every quantitative variable I could extract from ~7 years' worth of EMR data, I developed the conservative estimate that CDI costs Mount Sinai at least $1.5 million per year based on the attributable lengthening of inpatient stays and not including opportunity costs or the cost of isolation procedures. As an example of how to apply this estimate, from a purely economic perspective, this justifies investing at least $750,000 in any intervention that is projected to decrease CDI incidence by 50%.[7]

[6] Cooper et al. (2015), "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured".

[7] Previous studies suggest that interventions as simple as increased hand hygiene enforcement can achieve this magnitude of a drop; see Khanafer et al. (2015).

Finally, in Chapter 6, we applied a triad of extremely recent omics technolo-

gies to dissect the host response to chikungunya virus (CHIKV) in 42 pediatric patients from Managua, Nicaragua. Spread by the same mosquitos as the dengue and Zika viruses, CHIKV is an alphavirus that causes arthritic disease, sometimes lasting years, in millions of people around the world; in the US, local transmissions were first reported in 2013. The intent of this study was not only to capture a more "global" profile of the innate immune mechanisms underlying a poorly understood viral disease, but also to search for biomarkers for clinical outcomes that could lead to mechanistic insights and speed development of therapies, including vaccines. From a clinical perspective, this is the most futuristic of the projects in this thesis, as it will admittedly be some time before technologies like RNA-seq and Cytometry by Time-of-Flight (CyTOF) could be deployed in first-world hospitals, much less resource-poor countries like those most afflicted by mosquito-borne viruses. If reliable biomarkers for particular outcomes are revealed by these assays, however, they could be adapted into cheaper assay formats like lateral flow immunochromatography (the technology in over the counter pregnancy tests) or PCR.

From a research perspective, in Chapter 6 I was able to establish that monocytes seem to be the focus of both the phenotypic changes among peripheral blood monocytes (PBMCs) and the shifts in serum cytokines during the acute phase of infection. In particular, I found that "intermediate" $CD14^{++}CD16^{+}$ monocytes and an activated subpopulation of $CD14^{+}$ monocytes were most strongly associated with acute infection among all PBMC subtypes, and that monocytes display the greatest increases in CHIKV surface protein expression for the acute phase compared with convalescence. Along with a significant correlation discovered between acute phase "nonclassical" $CD14^{+}CD16^{++}$ monocytes and convalescent phase immunoglobulin G (IgG) titers, this was strongly reminiscent of a recently discovered mechanism for the innate immune response to dengue virus whereby CD16 upregulation in $CD14^{+}$ monocytes led to plasmablast differentiation and IgG secretion.[8] Establishment of a monocyte-centric reponse, along with our additional surprising discovery of substantial B cell expression of CHIKV surface protein, may redirect future investigations of leukocyte subpopulations that are critical for rapidly clearing CHIKV infection and potentially preventing chronic symptoms.

[8] Kwissa et al. (2014), "Dengue Virus Infection Induces Expansion of a CD14+ CD16+ Monocyte Population that Stimulates Plasmablast Differentiation".

Furthermore, I reported biomarkers that associated with specific clinical outcomes in Chapter 6. Among a transcriptome-wide analysis for differentially ex-

pressed transcripts associated with symptom severity and convalescent phase IgG titers, I noted the prominence of human leukocyte antigen (HLA) genes among the most significant results. This is mechanistically sensible, as these gene products are principally involved in antigen presentation and both the adaptive and cellular immune responses, and variants in HLA loci are well-known to correlate with changes in infection outcomes. More surprisingly, I found a globally significant association between expression of *MXRA7*, an essentially uncharacterized gene encoding a single-pass membrane, and acute phase symptom severity. Although too little is known about *MXRA7* to speculate on mechanisms, this result could spur further work on how this gene might be involved in innate immune responses, considering that it has only associated thus far with two other human disease processes in preliminary studies—namely, impaired retinal arterial microcirculation and endometriosis.[9]

To synthesize all the information created by the omic assays used in Chapter 6, I created the first multiscale network (to our knowledge) that associates both CyTOF and transcriptomic data with clinical outcomes. This analysis, presented in Figure 6.28, provides a "roadmap" for the array of immune responses I observed for CHIKV and how they correlate with the measured clinical variables. I verified that the dimensionality reductions caused by clustering both the CyTOF and RNA-seq data did not substantially degrade the model's ability to predict the major contrast of the study design, the timepoint of infection. Although this model has yet to be validated in targeted experiments on model organisms, I believe it already succeeds at compactly summarizing hundreds of thousands of changes in gene expression and cellular phenotypes. I also hope that the creation of similar multiscale networks for other viral infections—e.g., Dengue and Zika, as planned by the Dengue Human Immune Profiling Consortium that created the data for Chapter 6—will soon lend insight into the differential effects of those viruses on the human immune system. Since arboviral infections in tropical urban regions now frequently involve re-infection by differing serotypes of the same virus[10] and coinfection or concurrent infection by multiple viruses,[11] I anticipate that dissembling the multitude of combinatorial effects will require increasingly comprehensive immune profiling and network analysis.

[9] Sim et al. (2013), "Genetic Loci for Retinal Arteriolar Microcirculation"; Veiga-Castelli et al. (2010), "Genomic alterations detected by comparative genomic hybridization in ovarian endometriomas".

[10] OhAinle et al. (2011), "Dynamics of dengue disease severity determined by the interplay between viral genetics and serotype-specific immunity."

[11] Waggoner et al. (2016), "Viremia and Clinical Presentation in Nicaraguan Patients Infected with Zika Virus, Chikungunya Virus, and Dengue Virus."

Lessons learned and future directions

I NOW CONSIDER major conclusions that can be drawn from collectively interpreting these projects and the lessons learned during their execution, while considering promising directions for future work.

Using de novo *assemblies for pathogen surveillance*

In unison, chapters 2-4 show that is possible to create *de novo* assemblies in an efficient and complete enough manner that automated analyses of hospital pathogens become possible. Given sufficent throughput on a sequencing platform, which undeniably remains more expensive for third-generation long-read systems like the PacBio RS II compared to short-read platforms like the Illumina HiSeq or MiSeq, these analyses could even be performed with enough rapidity to affect clinical management.

The decision to use long-read sequencing and *de novo* assembly in preference to alignment of short reads is still pocketed with landmines, though, and not quite as automatable as one would need for cheap NGS-based pathogen surveillance at the $25/isolate cost envisioned by the Modernizing Medical Microbiology group at Oxford.[12] For instance, despite our best efforts to make fixes automatically, we must still perform manual curation to completely close the main chromosome for ~30% of sequenced genomes. From our burgeoning *de novo* assembly based investigations, I have so far dicovered that there is often more diversity (and types of diversity) than expected (Chapter 2), that there are certainly regions of assembled genomes that will complicate phylogenetic analysis like phage regions and plasmids (Chapter 3), and that some of those regions certainly have functional relevance, as antibiotic resistance or toxin loci are often found in repeat-heavy, mobile elements.[13] From a long-term research perspective, this certainly offers an exciting new opportunity to examine the behavior of those genetic elements among hospital strains. For the present-day use case of detecting and preventing transmissions, it is probably overkill compared to strain typing and calling variants against a standard reference for each type using short-read NGS.

To make full use of the information gained by discovering and cataloging structural variants, we will need many more well-annotated and complete ge-

[12] Köser et al. (2012); also see http://modmedmicro.nsms.ox.ac.uk/

[13] Knight et al. (2015), "Diversity and Evolution in the Genome of Clostridium difficile"; Casas et al. (2006), "Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California."; Chen et al. (2014), "Carbapenemase-producing Klebsiella pneumoniae: molecular and genetic decoding."

nomes, probably well beyond the hundreds that the PSP has sequenced so far. For the purposes of reintegrating structural variation into phylogenetic analysis, this database of genomes will be necessary to estimate individual rates of recombination for each of the many types of mobile elements that are observed in the hospital strains. Because most genomic surveys of bacterial populations have used alignment-based methods, the literature is currently much more effective at establishing rates of single-nucleotide mutations than measurements of expected rates of recombination or horizontal gene transfer for most species involved in HAIs.

Furthermore, for finding variant-phenotype associations, genome-wide association studies (GWAS) for bacteria provide an exciting new frontier that is only now feasible with the development of faster algorithms for finding pan-genomes across thousands of bacterial chromosomes.[14] They provide the benefit of a potentially unbiased method for discovering loci associated with clinically significant phenotypes, but come with the numerous pitfalls that afflict human GWAS studies, such as the possibility of cryptic population structures, unknown shifts in mutation rates across the evolutionary history, and errors in base-calling and phenotyping—along with entirely new idiosyncrasies of having an entire genome under linkage disequilibrium and clonal strata in the input population.[15] As these methods mature, however, it will be exciting to apply them to the PSP dataset to discover bacterial genomic variants that correlate with increased virulence, transmission rates, or antimicrobial resistance.

Even for the relatively simple purpose of using NGS data for transmission detection, *de novo* assembly has introduced some wrinkles that still remain to be resolved. As mentioned above, using structural variants during assessments of evolutionary distance is currently much more complicated than looking at SNVs, so for the most part we have simply discarded structural variants during phylogenetic analyses. Furthermore, while using MUMmer[16] to find SNVs between two strains is incredibly fast, requiring only minutes per comparison, we still have yet to achieve reliable filtering of the SNV calls from regions that are expected to be hypervariable, like phage regions (Chapter 3). So far, our approach has been to accept these SNVs as expected "noise" and to repeat comparisons across all pairs of genomes to increase confidence in defining a cluster (heatmap analyses in Chapter 4). A different solution is to use new methods like Parsnp[17] to find a core genome for all same-species isolates, although this

[14] Brynildsrud et al. (2016), "Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary".

[15] Read and Massey (2014), "Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology"; Brynildsrud et al. (2016).

[16] Delcher, Salzberg, and Phillippy (2003), "Using MUMmer to identify similar regions in large sequence sets."

[17] Treangen et al. (2014), "The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes."

incurs the tradeoff of completely discarding structurally variable regions and cannot include genomes beyond a certain MUMi distance.[18] Yet another alternative is ClonalFrameML,[19] which can incorporate recombinatorial events into phylogenetic analysis, but does not include an alignment algorithm and is therefore dependent on an LCB-based aligner like progressiveMauve,[20] which cannot scale beyond small numbers of genomes. We are currently in the process of comparing distributions of Parsnp and MUMmer SNVs to see which will best serve the use cases of finding transmissions and reconstructing phylogenies going forward, and are already using PathogenDB-viz to visualize both types of distances.

[18] This is a distance metric constructed from counting shared maximally unique matches; see Deloger, El Karoui, and Petit (2009).

[19] Didelot and Wilson (2015), "ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes".

[20] Darling, Mau, and Perna (2010), "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement".

All told, the idea of a set of machines that can perform the workflow in Figure 1.2 without any human intervention remains fantastical, just as it is equally difficult to imagine automating the work of infection control officers or infectious diseases specialists. Computers are adept at accelerating the repetitive, algorithmic parts of the workflow but cannot foreseeably perform unguided inference at the level needed to, e.g., intelligently assess all potential sources of an ongoing outbreak. Even the best machine learning algorithms are hampered by the sources and quality of training data made available to them. As our experience shows, setting up these training data involves dedicated effort by expert human curators, who can eventually engineer automated methods for streaming these data into structured databases,[21] but they will always be needed again when the data violates an assumption of the automation—an eternal threat, given how messy clinical informatics data can be. Nevertheless, the level of automation that we have already achieved is sufficient for detecting outbreaks that merit intervention by infection control staff (Chapter 4), which is a baseline of utility that does fulfill an original goal of the PSP.

[21] An example is how we initially imported nightly reports from the EMR manually into PathogenDB, and then Harm van Bakel eventually automated this with a series of Perl scripts.

How do we best make use of genomic surveillance data?

Even if we can detect outbreaks well, this raises the spectre of how to best apply the data that would be produced by a theoretically optimal genomic surveillance system for hospital pathogens, as was pursued in Chapters 1–4. Additionally, such a system might be targeted to any particular subset of the specimens received by the clinical microbiology lab, along with proactive sampling of the environment, healthcare workers, and colonized patients, and none of the pre-

vious information really answers what extent of NGS surveillance is warranted. If the cost of hypothetically comprehensive NGS surveillance far outweighs the cost of HAIs and the potential benefits of preventing them, it would be unreasonable to expect NGS to ever be routinely used in this way. Chapter 5 attempted to lay down one cornerstone of a comprehensive answer to this question, namely the local cost of one HAI (giving an anchorpoint for a proportional investment in prevention), but there is clearly much more work to be done.

Conventional wisdom says that an ounce of prevention is worth a pound of cure. This is not exactly how hospital budgeting responds to the current incentives of the US healthcare system, which in general rewards hospitals for overconsumption and ordering expensive procedures for patients, even if those procedures incur the increased risk of HAIs.[22] Although there are some financial penalties for "causing" infections, such as the 2013 decision of Centers for Medicare and Medicaid Services (CMS) to penalize reimbursements to hospitals with high rates of *C. difficile* infection beyond two to three days of admission (which might be assumed to be caused by the healthcare provider),[23] these penalties likely will never stack up against the revenue incentives of ordering the procedures that may have led to those HAIs. Furthermore, the chain of causality is much more complicated than assumed by such a simplistic incentive; e.g., it's clear from our data that a lot of *C. difficile* cases are not caused by patient-to-patient transmission within the hospital, as seen in NGS studies at other hospitals.[24] If patients present with pre-existing colonization by *C. difficile*, it's less clear whether a microbiotic dysregulation during the patient stay that leads to symptomatic infection (even if caused by a hospital-administered antibiotic) is the "fault" of the healthcare provider. The typical response to the new CMS incentive is for hospitals to prioritize *C. difficile* testing for any patient presenting with diarrhea, as it is desirable to prove the patient was pre-colonized by a positive result before the third day passes so that a later positive result does not count against the hospital. In a sense, this only encourages a different kind of overconsumption.

Even if a hypothetically optimal NGS surveillance system could completely disprove patient to patient transmissions were occurring, could this data really be used to create more well-aligned financial incentives? NGS of patient specimens cannot rule out that HAIs are coming from healthcare staff, unless they are swabbed as well, and several times per day—probably not a reasonable expec-

[22] Lallemand (2012), "Health Policy Brief: Reducing Waste in Health Care"; Gawande (2015), "Overkill".

[23] Drozd et al. (2015), "Mortality, hospital costs, payments, and readmissions associated with Clostridium difficile infection among Medicare beneficiaries".

[24] Eyre et al. (2013), "Diverse sources of C. difficile infection identified on whole-genome sequencing."; Didelot et al. (2012), "Microevolutionary analysis of Clostridium difficile genomes to investigate transmission."; Roach et al. (2015), "A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota".

tation. Another pernicious source would be the hospital environment. Many studies show that rates of CDI go down after a deep clean of patient rooms,[25] although these studies typically cannot prove that subsequent reduced CDI rates were primarily attributable to reduction of environmental sources.[26] Without actively testing all patients on admission, there is also no way to prove that a patient was colonized with *C. difficile* and underwent microbiotic dysregulation. Therefore, this all points toward NGS only providing suggestive—not definitive—proof of the source of all cases of CDI even in the best of circumstances, and therefore, even if CMS or other payers would ever consider evidence from genomic analyses, it seems unlikely that any level of surveillance could perfectly re-align economic incentives so that market forces "solve" CDI.

This may be disheartening to economists, but to take the bold step of setting aside economic arguments, hospitals still have an ethical obligation to take reasonable precautions to prevent HAIs under a Hippocratic duty of "doing no harm."[27] In this context, genomic surveillance at the level described in Chapter 4, even if not quite in real time and limited to finding cases of patient-to-patient transmission, is still immensely valuable in generating a crisp metric for instances of system-level failure that deserve attention. In the absence of such a metric, choosing when and where to deploy infection control interventions is depressingly arbitrary. Hospital epidemiologists currently judge danger levels based on epidemiological data, whose natural variation can create the appearance of an outbreak when none exists. In fact, two of the first investigations conducted by the PSP (one of which is Chapter 2) were in response to perceived outbreaks that were revealed to be random temporal clusters by NGS data.[28] Proven patient-to-patient transmissions, on the other hand, are a much stronger reason for alarm, as there are few circumstances that could cause patients to have clonal positive cultures that do not require some failure of standard infection control practices, like equipment sterilization or hand hygiene.[29] Since proven patient-to-patient outbreaks are perhaps the worst kind of disaster that infection control officers are expected to prevent at all costs, NGS surveillance is undoubtedly useful for being able to sound the loudest, earliest alarm that preventable harm is occurring and may be on the cusp of harming more patients. As previously stated, if aiming for this goal alone, as of today short read NGS is likely simpler, faster, cheaper, and just as effective as the methods used in this dissertation, although this will likely change with continuing im-

[25] Best et al. (2014), "Effectiveness of deep cleaning followed by hydrogen peroxide decontamination during high Clostridium difficile infection incidence"; Hughes et al. (2013), "Impact of cleaning and other interventions on the reduction of hospital-acquired Clostridium difficile infections in two hospitals in England assessed using a breakpoint model"; Khanafer et al. (2015).

[26] For example, staff may be more conscious about hand hygiene after seeing their unit being cleaned, which could only be controlled by a "sham clean", which was not performed in any of the studies. Revamping hand hygiene protocols is equally linked to changes in CDI rates; see Khanafer et al. (2015).

[27] Less nobly, there is still tort liability.

[28] For other examples, see Peaper et al. (2015) and Anderson et al. (2014).

[29] One of the potential reasons would be a common source of contamination in the clinical microbiology lab or during specimen collection, but this is also valid cause for alarm.

provement of long-read technologies and associated computational tools.

Other longer term applications of NGS may be able to answer previously intractable mysteries for health systems that are grappling with multidrug-resistant HAIs, which is essentially every health system in the world. Increasing antimicrobial resistance is a serious global problem, and the oft-touted strategy of decreasing antibiotic prescriptions has not yet been proven to reverse this trend or improve outcomes.[30] Chapter 2 highlighted the power of using completely assembled genomes to reveal the genetic mutations occurring within hospital strains that develop resistance during antimicrobial therapy. A preliminary assessment of 317,240 blood cultures from EMR data between 2009 and 2014 at Mount Sinai found that 875 visits included serial cultures where the same species was isolated more than once, among which 175 susceptible-to-resistant transitions for a particular antibiotic–species combination were observed. Of these, 91 involved *Staphylococcus epidermidis* and therefore likely were contaminated samples, but I also observed 44 transitions for *Klebsiella pneumoniae*, 27 for *Pseudomonoas aeruginosa*, 19 for *Escherichia coli*, and 17 for *Staphylococcus aureus*. A future study could use these instances plus the existing sequencing capabilities of the PSP to survey the genetic causes behind all instances of emerging resistance in the hospital, which is convincingly attributable to hospital treatment when concurrent admistration of the same class of antibiotics is recorded by the EMR (see Figure 7.1). Then, following in the footsteps of a recent study showing that fluoroquinolone restriction associated with decreases in fluoroquinolone-resistant *C. difficile* incidence and NGS-detected transmissions,[31] the PSP could precisely examine the impact of antimicrobial stewardship programs at Mount Sinai on the molecular evolution of local HAI strains. These future directions are currently being pursued by colleagues in the Bakel lab.

[30] Infectious Diseases Society of America (2010), "The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020."; Wagner et al. (2014), "Antimicrobial stewardship programs in inpatient hospital settings: a systematic review."

[31] Dingle et al. (2017), "Effects of control interventions on Clostridium difficile infection in England: an observational study."

The need for better clinical informatics

Ultimately, as also hinted at by Chapter 5, measuring the impact of interventions at the hospital level continues to be constrained by the quality, accessibility, and depth of clinical informatics data. One could argue that the least-developed part of the vision originally introduced in Figure 1.2, despite all of the progress evidenced in Chapters 2-4, is the development of comprehensive "Management Strategies" as a result of our analyses. I could certainly report

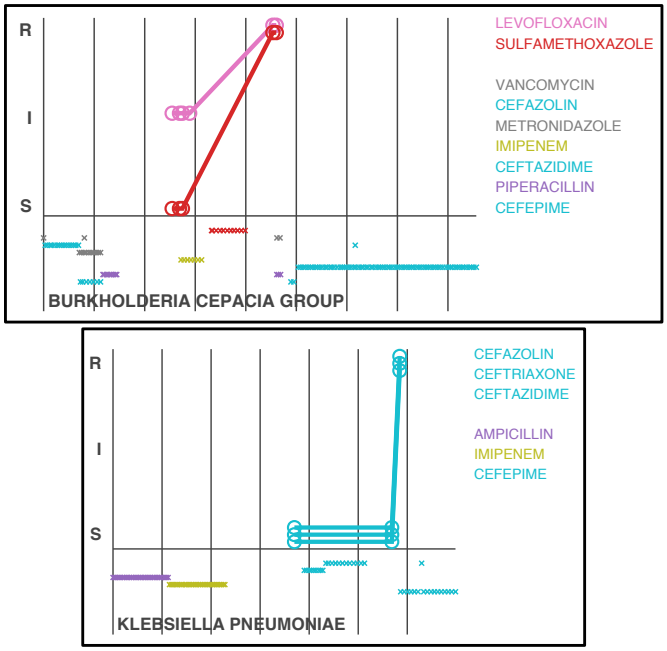


Figure 7.1: **Cases of emerging resistance mined from electronic medical records.** Records of 317,240 blood cultures performed at Mount Sinai between 2009 and 2014 were mined for same-species isolates captured during a contiguous hospital visit that showed at least one susceptible-to-resistant transition in Vitek (bioMeríeaux) drug susceptibility test (DST) results for at least one antibiotic. Two representative examples are plotted here. Time is displayed on the X axes, with each vertical line representing one week and $t = 0$ representing the admission timestamp. DST results for the antibiotics involved in at least one transition are plotted as the circles above the X axis. Recorded administrations of antibiotics, colored by antibiotic class, are marked with crosses below the X axis, and vertically ordered using the order of the legend. Note that the transition time periods correspond to concurrent administrations of antibiotic(s) of the same class. S, susceptible; I, indeterminate; R, resistant.

transmissions along with other associations in our data, but justifying any specific response based on this information is an open question. For example, despite my best efforts to model *C. difficile* risk among variables in the EMR, I quickly faced dwindling statistical power and interpretability when trying to reformat modeled risk parameters into concrete practice recommendations. This is because EMR data is observational, the hierarchy of entered data is subject to many varieties of bias, and every variable has dozens of hidden correlates. For example, even if administration of promethazine is a top ranking correlate for CDI at Mount Sinai according to our models (and it was, for four out of five CDI definitions) it would be somewhat ridiculous to suggest that banning use of promethazine, a preoperative sedative, will decrease rates of CDI.[32]

Furthermore, I was only able to estimate a correlate for the cost of CDI: the changed length of stay, not a true dollar figure. This is certainly better than no figure at all or a regional estimate, but it is easy to imagine individual situations in which differences in the length of stay do not correlate well with true costs, either because of expensive procedures resulting from CDI (like a colectomy) or downstream effects at future visits to hospitals. The pursuit of better cost data might seem dogged and perhaps foolhardy, but as US healthcare providers incur more of the financial risk of managing populations[33] it is impossible to expect them to rationally manage resources without attempting to understand

[32] Our best guess for why this variable ranks so highly is that it correlates with surgeries, and many surgical procedures do in fact increase the risk of CDI by incurring rounds of antibiotic prophylaxis and increased exposure to the hospital environment.

[33] Shortell et al. (2015), "Accountable Care Organizations: The National Landscape."

the true costs of care at an individual level.[34] This flies in the face of most current hospital billing practices, which often purposefully obscure the true costs of care in order to set high prices that maximize payout from insurers.[35] An unfortunate side effect is that diagnosis and procedure codes created during the billing mechanism are likely skewed toward values that incur higher reimbursement.[36] Unless these practices change, the ability to estimate per-patient costs will continue to suffer.

Therefore, integrating cost and clinical informatics data into models that allow infection control staff to analytically propose management strategies remains a challenging endeavor. Even operating on EMR data alone, as I did in Chapter 5, required clearance of substantial technological and bureaucratic hurdles. Although technology that supports data warehouses for EMR and clinical microbiology data has been available for some time,[37] in practice, making best use of these data requires a substantial amount of validation, curation, and indexing that no vendor will ever be able to provide in an "out-of-box" solution, as every practice environment has its unique operational quirks. In our case, I encountered irregularities in the data resulting from overaggressive deidentification procedures, shifting report formats due to an upgrade of the Vitek software, and many duplicated lab results that needed to be merged, among other issues. All of these required custom software engineering before I could begin analysis. On top of this, most infection control analyses (like those performed by the PSP) will need a partially-identified dataset, since epidemiological assessment of outbreaks requires real locations and times and these are considered identifiers for protected health information according to the HIPAA Privacy Rule.[38] This complicates institutional approvals for accessing large EMR datasets, and depending on the warehousing and deidentification techniques used, it may even make it impossible to associate the EMR data with specimens that are sent to the clinical microbiology lab.

Should a team like the PSP become increasingly efficient at integrating new data sources and training new statistical models on the entire dataset, the notion of a continuously learning healthcare system for infectious diseases is still within reach. It remains constrained by the dangers of inferring causality from observational data,[39] since it is very unlikely that a hospital would be able to perform randomized clinical trials for all of its infection control interventions (and it would certainly be unethical to randomize exposure of patients to known

[34] Hilsenrath, Eakin, and Fischer (2015), "Price-transparency and cost accounting: Challenges for health care organizations in the consumer-driven era".

[35] Ibid.

[36] Rhee et al. (2015), "Improving documentation and coding for acute organ dysfunction biases estimates of changing sepsis severity and burden: a retrospective study"; Romano and Mark (1994), "Bias in the coding of hospital discharge data and its implications for quality assessment."

[37] Isniewski et al. (2003), "Development of a Clinical Data Warehouse for Hospital Infection Control".

[38] See 45 CFR §164.514(b)(2).

[39] Dahabreh and Kent (2014), "Can the learning health care system be educated with observational data?"

HAI sources, which would seem to be necessary to definitively prove that a preventative measure works). However, to the extent that comprehensive models can be constructed from EMR data (Chapter 5) that are monitored before and after interventions are deployed, the PSP could determine to the best possible extent how much of an impact each intervention has on HAI rates and outcomes and rationally deploy these interventions in response to early-warning systems for detecting patient-to-patient transmissions in real time (Chapter 4). There is no doubt that this will require substantial investment at the health system level, given the ongoing costs of building and maintaining the infrastructure for all of these data sources.

New technologies for profiling the host response to infection

For over a hundred years, clinical microbiology techniques have focused almost exclusively on culturing and phenotyping the pathogen and not the host.[40] While studies of cancer and neurodegenerative diseases have pushed aggressively toward using omic techniques like RNA-seq to profile the host for diagnostic and prognostic purposes,[41] corresponding uses for infectious diseases have been slower to emerge, perhaps because the pathogens themselves have been such conducive targets. However, given our natural ability to trace infectious diseases to a root perturbation of the host—namely the introduction of a specific pathogen—it only seems logical to apply similar omic assays on hosts at proceeding timepoints to reveal biomarkers for the various outcomes of that perturbation and create the capacity to predict them. In diseases where there are a wide range of outcomes, such as viral diseases where most infections are asymptomatic but a rare few are eventually lethal, this could have direct prognostic value. For infections where the innate immune reponse is not well understood, these methods could eventually lead to a better understanding of the mechanisms underlying better and worse outcomes.

[40] Didelot et al. (2012), "Transforming clinical microbiology with bacterial genome sequencing".

[41] Costa et al. (2012), "RNA-Seq and human complex diseases: recent accomplishments and future perspectives".

Chapter 6 used three omic techniques to comprehensively profile the host response to CHIKV infection. From a research perspective, this study is valuable simply for providing the most global look at molecular and cellular perturbations induced by the virus, as it produced new signatures at the cellular and serum cytokine levels that consistently centered on monocytes, found that B cells expressed globally significant levels of CHIKV surface protein during infection despite no prior evidence for B cell tropism, and identified specific

gene modules that associate with acute infection and higher viremic load. It also valiantly attempted to synthesize this maelstrom of changes into a more intepretable picture (Figure 6.28).

The study also attempted to tie omic changes to clinical variables like convalescent IgG titers and acute phase symptom severity, although the signatures for these changes were constrained to one assocation among CyTOF subpopulations and two small transcriptomic signatures. These signatures merit reinvestigation in a large cohort to assess their reproducibility. In developing these signatures, I was constrained by the cohort design, which tried to maximize the acute-convalescent contrast while minimizing other contrasts, since the acute-convalescent signature was considered most important. In hindsight, given the overwhelming strength of the observed signature for timepoint, the study may have benefitted from a cohort that had more variation in disease severity if a substantial signature for that outcome variable were desired. Most interestingly, once follow-up data on these cases become available, if chronic symptoms develop in any of the patients (at the time of the study, they had not) it would be worthwhile to search for a new signature for this particularly worrisome outcome. Previous profiling studies followed cohorts for a year or more to identify some potential biomarkers for chronic post-CHIKV arthralgia,[42] but given the extent of our dataset, globally significant markers could be determined with much more confidence. Follow-up samples for a far off timepoint (like one year post-infection) may also establish a baseline picture that mitigates potential criticism that the convalescent samples were not representative of a truly "healthy" state, although symptoms and viremia had resolved for all patients in our particular cohort.

Some of the other results would benefit from validation in more focused experiments. While the B cell expression of CHIKV surface protein is suggestive of viral entry and replication, given that B cells have not yet been found to be productively infectable by CHIKV in vitro,[43] it would help to reassess some samples with antibodies for both CHIKV surface and nonstructural proteins.[44] The subpopulation of $CD14^+$ monocytes that associated with acute infection but displayed a previously unreported phenotype (positivity of CCR4, CXCR3 and CCR6, which are better associated with T cells) would benefit from isolation and characterization in vitro via standard flow cytometry, both to confirm that this phenotype is real and to see whether it can be directly induced from in-

[42] Poo et al. (2014), "Multiple Immune Factors Are Involved in Controlling Acute and Chronic Chikungunya Virus Infection"; Chaaitanya et al. (2011), "Role of proinflammatory cytokines and chemokines in chronic arthropathy in CHIKV infection."; Hoarau et al. (2010), "Persistent chronic inflammation and infection by Chikungunya arthritogenic alphavirus in spite of a robust host immune response."; Schilte et al. (2013), "Chikungunya Virus-associated Long-term Arthralgia: A 36-month Prospective Longitudinal Study".

[43] Her et al. (2010), "Active infection of human blood monocytes by Chikungunya virus triggers an innate immune response"; Sourisseau et al. (2007), "Characterization of reemerging chikungunya virus."; Teng et al. (2012), "Viperin restricts chikungunya virus replication and pathology".

[44] Concurrent expression of nonstructural *and* surface proteins is better evidence of entry and active replication, rather than surface proteins alone, which may only reflect virions bound to the outside of cells.

active $CD14^+$ monocytes after inoculation with CHIKV. The association of several HLA transcripts with clinical variables was mechanistically plausible, but the strong association between the little-understood gene *MXRA7* and symptom severity was surprising. Given that I used a very simple rubric for severity in a relatively homogenous cohort, this result should be validated by either RNA-seq or RT-PCR on samples from a larger, more heterogeneous cohort. If the result holds up, it would be interesting to assess gene expression of *MXRA7* among leukocyte subpopulations during in vitro infection, so as to start narrowing down what its role in the innate immune response might be.

All of these experiments are plausible lines of future inquiry by the Dengue Human Immune Profiling Consortium, which also plans to develop comprehensive immune profiles for other arboviral infections. Once this profiling is complete for all ~200 biosamples (in approximately three years), a unified network model for all three diseases will be attempted, and the sample size will also then be sufficient for fitting Bayesian models on coexpression modules that strongly associate with the various clinical outcomes. Bayesian models permit the inference of causality between differentially expressed genes and the identification of key drivers for the innate immune response. In short, I believe that the potential applications of this dataset and the involved immune profiling techniques have only just begun to be explored.

Conclusions

The introductory chapter of this dissertation claimed, ambitiously and with only extrapolations as evidence, that "a potent combination of NGS and EMR data will transform infectious disease management." Although this thesis has not singlehandedly completed that transformation, I find no reason to change this prediction, given the substantial progress that was demonstrated here for a few short years of effort. As I had expected, the unprecedented specificity of the PSP's NGS data allows us to reconstruct and detect clandestine transmissions of HAIs within The Mount Sinai Hospital, and large parts of this process are now automated and easily communicated to clinicians with the software that we have built. The use of long read NGS and *de novo* assembly to support pathogen surveillance is novel—and moreover, looks to be essential for characterizing the recombination and horizontally transferred genes that underlie

evolution of virulence and antibiotic resistance in HAI organisms, particularly given the diversity of many gram negatives. As the PSP's linked dataset of genomes and clinical metadata grows, I still fully expect that the PSP can develop predictive models for virulence, resistance, and infection outcomes based on comprehensive, pan-genomic analyses. Given continued investment in the clinical informatics available to the PSP, it will be able to establish the local costs of other HAIs, measure effects correlated with various interventions, and support specific recommendations for new management strategies. Meanwhile, technologies available for profiling the host response to infection continue to grow, and my study of chikungunya virus indicates that omic assays can simultaneously establish globally significant trends in the immune response (like the dominance of monocyte involvement) and extremely subtle biomarkers for clinical outcomes, like infection severity.

In conclusion, this dissertation found that integrative analysis of EMR, NGS, and other omics data was an fruitful strategy for addressing critical components of several urgent problems in infectious diseases, such as the spread of HAIs, increasing rates of antimicrobial resistance, and comprehensive profiling of the host response to a newly epidemic, poorly understood virus. Therefore, I believe that multiscale analysis is indeed transforming clinical infectious diseases.

A

Appendix Tables

I know I've made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal. I've still got the greatest enthusiasm and confidence in the mission. And I want to help you.

—HAL 9000, *2001: A Space Odyssey*

TARS: [as Cooper repairs him] *Settings. General settings. Security settings.*

COOPER: *Honesty, new setting: ninety-five percent.*

TARS: *Confirmed. Additional settings.*

COOPER: *Humor, seventy-five percent.*

TARS: *Confirmed. Self destruct sequence in T minus 10, 9...*

COOPER: *Let's make that sixty percent.*

TARS: *Sixty percent, confirmed. Knock knock.*

COOPER: *You want fifty-five?*

—*Interstellar*

Table A.1: **Variables closely correlated with *Clostridium difficile* infection workup or treatment that were excluded from propensity modeling in Chapter 5.** Raw data for this table are available in tab-separated values format from Figshare at DOI: 10.6084/m9.figshare.4311695. Abbreviation: C.DIFF, *Clostridium difficile*; PCR, polymerase chain reaction; EIA, enzyme immunoassay; CAP, caplet; TAB, tablet; ISO-OSM, iso-osmotic; IV, intravenous; SUSP, suspension

Variable	Description
problem_list:008.45	Intestinal infection due to Clostridium difficile
abnormal_labs:4730	C.DIFF. TOXIN B PCR
abnormal_labs:4647	C.DIFF EIA
abnormal_labs:4647	C.DIFF EIA TOXIN A&B
meds_administered:300025	VANCOMYCIN ORAL LIQUID REPACKAGE ONLY - 125MG/2.5ML
meds_administered:63653	VANCOMYCIN ORAL
meds_administered:8200	VANCOMYCIN 125 MG CAP
meds_administered:14246	VANCOMYCIN 250 MG/5 ML ORAL SOLUTION
meds_administered:300092	VANCOMYCIN ENEMA
meds_administered:13623	VANCOMYCIN 250 MG CAP
meds_administered:5484	METRONIDAZOLE 250 MG TAB
meds_administered:54227	METRONIDAZOLE ORAL
meds_administered:6870	METRONIDAZOLE 500 MG TAB
meds_administered:5484	METRONIDAZOLE 250 MG TAB
meds_administered:6295	METRONIDAZOLE IN SODIUM CHLORIDE (ISO-OSM) 500 MG/100 ML IV PIGGY BACK
meds_administered:400648	METRONIDAZOLE 250 MG/50 ML ISO-OSMOTIC SOLUTION
meds_administered:1527	METRONIDAZOLE HCL 500 MG IV SOLUTION
meds_administered:19702	METRONIDAZOLE 375 MG CAP
meds_administered:300009	METRONIDAZOLE 50 MG/ML ORAL SUSP
meds_administered:400660	METRONIDAZOLE 750 MG/150 ML ISO-OSMOTIC SOLUTION
meds_reported:300025	VANCOMYCIN ORAL LIQUID REPACKAGE ONLY - 125MG/2.5ML
meds_administered:14246	VANCOMYCIN 250 MG/5 ML ORAL SOLUTION
meds_administered:300092	VANCOMYCIN ENEMA
meds_administered:13623	VANCOMYCIN 250 MG CAP
meds_administered:5484	METRONIDAZOLE 250 MG TAB
meds_administered:54227	METRONIDAZOLE ORAL
meds_administered:6870	METRONIDAZOLE 500 MG TAB
meds_administered:5484	METRONIDAZOLE 250 MG TAB
meds_administered:6295	METRONIDAZOLE IN SODIUM CHLORIDE (ISO-OSM) 500 MG/100 ML IV PIGGY BACK
meds_administered:400648	METRONIDAZOLE 250 MG/50 ML ISO-OSMOTIC SOLUTION
meds_administered:1527	METRONIDAZOLE HCL 500 MG IV SOLUTION
meds_administered:19702	METRONIDAZOLE 375 MG CAP
meds_administered:300009	METRONIDAZOLE 50 MG/ML ORAL SUSP
meds_administered:400660	METRONIDAZOLE 750 MG/150 ML ISO-OSMOTIC SOLUTION
meds_reported:300025	VANCOMYCIN ORAL LIQUID REPACKAGE ONLY - 125MG/2.5ML
meds_reported:63653	VANCOMYCIN ORAL
meds_reported:8200	VANCOMYCIN 125 MG CAP

Continued on next page…

Variable	Description
meds_reported:14246	VANCOMYCIN 250 MG/5 ML ORAL SOLUTION
meds_reported:300092	VANCOMYCIN ENEMA
meds_reported:13623	VANCOMYCIN 250 MG CAP
meds_reported:5484	METRONIDAZOLE 250 MG TAB
meds_reported:54227	METRONIDAZOLE ORAL
meds_reported:6870	METRONIDAZOLE 500 MG TAB
meds_reported:5484	METRONIDAZOLE 250 MG TAB
meds_reported:6295	METRONIDAZOLE IN SODIUM CHLORIDE (ISO-OSM) 500 MG/100 ML IV PIGGY BACK
meds_reported:400648	METRONIDAZOLE 250 MG/50 ML ISO-OSMOTIC SOLUTION
meds_reported:1527	METRONIDAZOLE HCL 500 MG IV SOLUTION
meds_reported:19702	METRONIDAZOLE 375 MG CAP
meds_reported:300009	METRONIDAZOLE 50 MG/ML ORAL SUSP
meds_reported:400660	METRONIDAZOLE 750 MG/150 ML ISO-OSMOTIC SOLUTION
problem_list:041.84	Other specified bacterial infections in conditions classified elsewhere and of unspecified site, other anaerobes
problem_list:V07.0	Need for isolation
problem_list:V02.3	Carrier or suspected carrier of other gastrointestinal pathogens
problem_list:787.91	Diarrhea

Table A.2: **Antibodies used for CyTOF analysis in Chapter 6.** All antibodies were either purchased directly from Fluidigm or conjugated in-house using X8 conjugation kits.

Isotope	Target	Clone	Source
113 In	CD57	HCD57	Biolegend
115 In	CD45	HI130	Biolegend
142 Nd	CD19	HIB19	Biolegend
143 Nd	CD45RA	HI100	Biolegend
144 Nd	CD141	M80	Biolegend
145 Nd	CD4	RPA-T4	Biolegend
146 Nd	CD8	RPA-T8	Biolegend
147 Sm	CD20	2H7	Fluidigm
148 Nd	CD16	3G8	Biolegend
149 Sm	CD127	A019D5	Biolegend
150 Nd	CD1c	L161	Biolegend
151 Eu	CD123	6H6	Biolegend
152 Sm	CD66b	G10F5	Biolegend
153 Eu	CXCR5	RF8B2	Fluidigm
154 Sm	CD86	IT2.2	Biolegend
155 Gd	CD27	O323	Biolegend
156 Gd	CCR5	NP-6G4	Fluidigm
158 Gd	CHIKV	CHK-152	Biomatik
159 Tb	CD11c	Bu15	Fluidigm
160 Gd	CD14	M5E2	Biolegend
161 Dy	CD56	B159	BD Biosciences
162 Dy	CD80	2D10.4	Fluidigm
163 Dy	CCR4	205410	R&D Systems
164 Dy	CD40	5C3	Biolegend
165 Ho	CCR6	G034E3	Biolegend
166 Er	CD25	M-A251	Biolegend
167 Er	CCR7	G043H7	Biolegend
168 Er	CD3	UCHT1	Biolegend
169 Tm	CX3CR1	2A9-1	Biolegend
170 Er	CD38	HB-7	Biolegend
171 Yb	CD161	HP-3G10	Biolegend
172 Yb	CD209	9E9A8	Biolegend
173 Yb	CXCR3	G025H7	Biolegend
174 Yb	HLADR	L243	Biolegend
175 Lu	PD-1	EH12.2H7	Fluidigm
176 Yb	CD54	HCD54	Biolegend
209 Bi	CD11b	ICRF44	Fluidigm
194Pt	CD45 (acute BC)	HI130	Biolegend
198Pt	CD45 (conv BC)	HI130	Biolegend
103 Rh	Viability	n/a	Fluidigm
191/193 Ir	DNA	n/a	Fluidigm

B

Appendix Figures

Hold on. You have to slow down. You're losing it. You have to take a breath. Listen to yourself. You're connecting a computer bug I had with a computer bug you might have had and some religious hogwash. You want to find the number 216 in the world, you will be able to find it everywhere. 216 steps from a mere street corner to your front door. 216 seconds you spend riding on the elevator. When your mind becomes obsessed with anything, you will filter everything else out and find that thing everywhere.

—SOL ROBESON, *Pi*

Calm down, Doctor! Now's not the time for fear. That comes later.

—BANE, *The Dark Knight Rises*

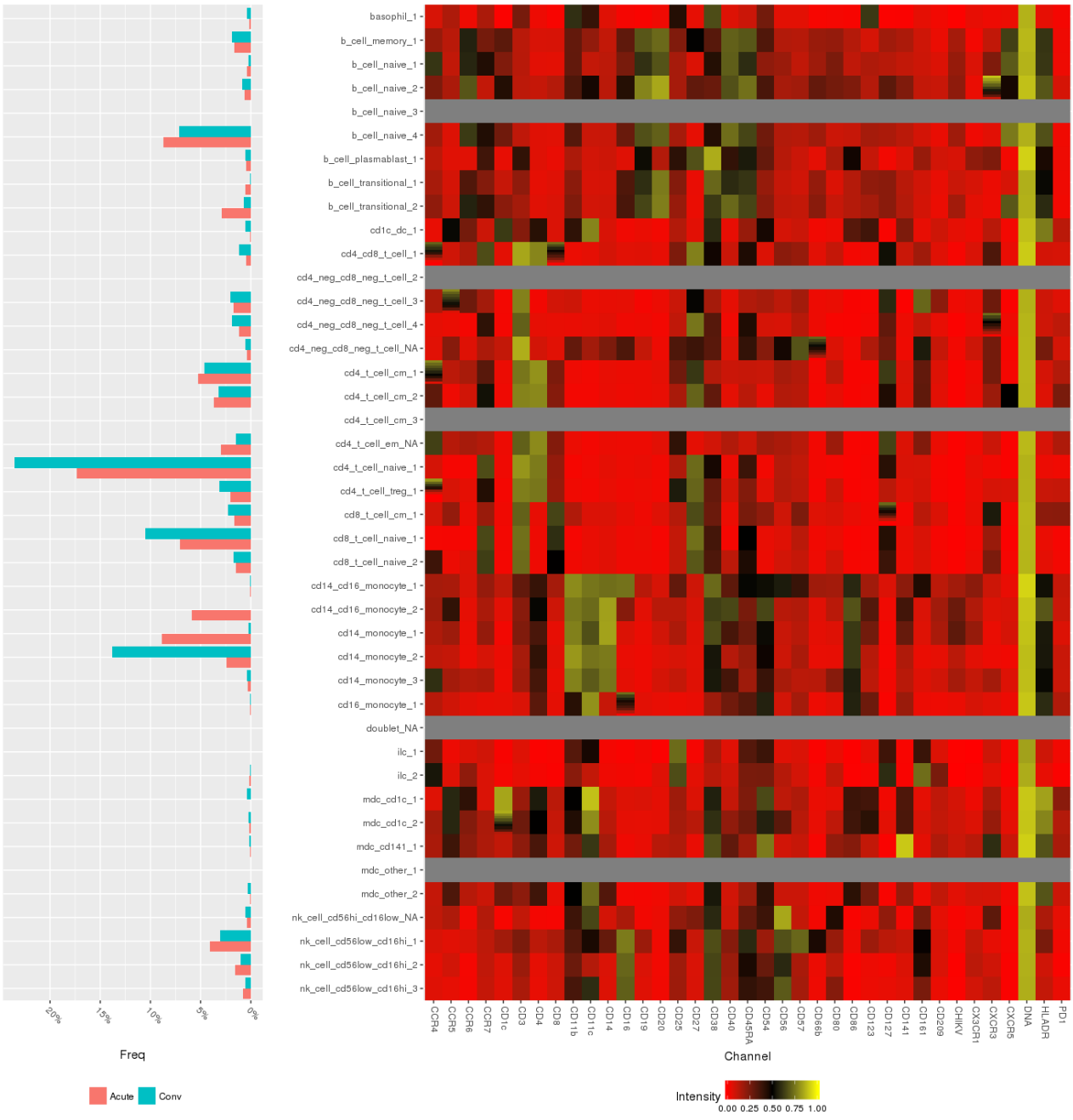


Figure B.1: **Example output of `MetaHybridLouvain`** for the representative sample used in Figure 6.2, Figure 6.6, and 6.8. At left, frequencies for each sub-community at each timepoint, and at right, mini-heatmaps of channel values for each sub-community (plotted within each row). As expected, sub-communities generally show similar values among all channels for their constituent events, with some exceptions (vertical gradients within mini-heatmaps). *Continued in Figure B.2.*

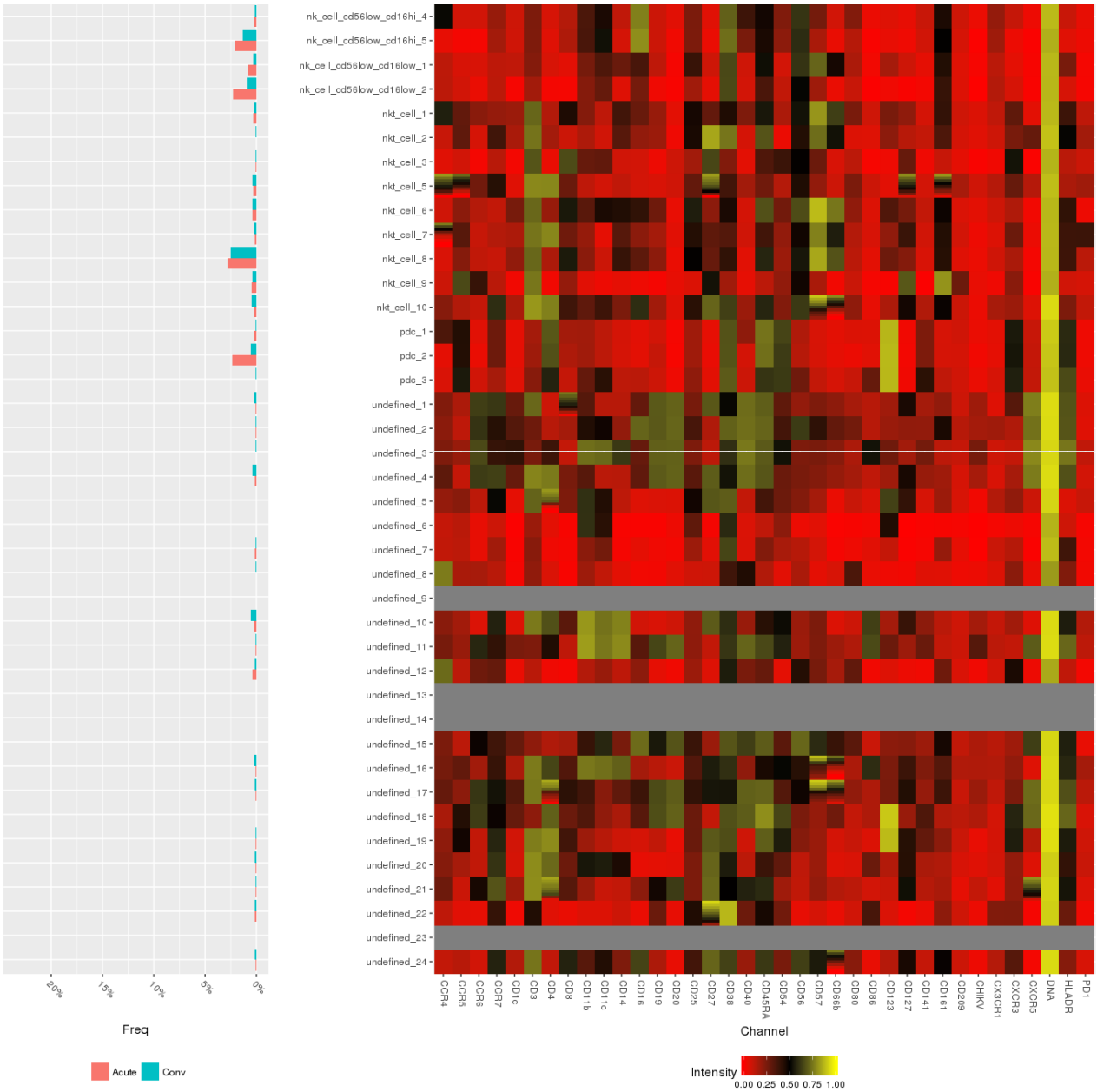


Figure B.2: **Continuation of example output of `MetaHybridLouvain`** for the representative sample used in Figure 6.2, Figure 6.6, and 6.8. At left, frequencies for each sub-community at each timepoint, and at right, mini-heatmaps of channel values for each sub-community (plotted within each row). As expected, sub-communities generally show similar values among all channels for their constituent events, with some exceptions (vertical gradients within mini-heatmaps). *Continued from Figure B.1.*

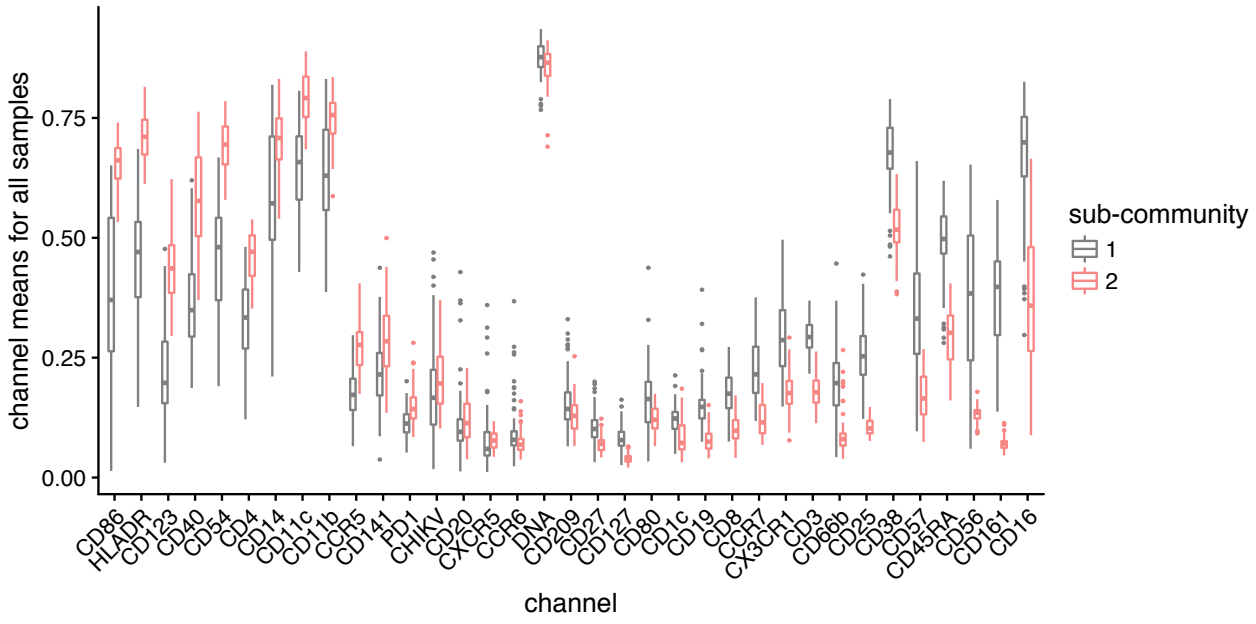


Figure B.3: **Differences in per-sample channel means between two CD14^{+}CD16^{+} sub-communities identified by `MetaHybridLouvain`.** 1 corresponds with the "intermediate" CD14^{++}CD16^{+} phenotype, while 2 corresponds with the "nonclassical" CD14^{+}CD16^{++} phenotype. The X axis is filtered to only the channels with differences significant at FDR < 0.05, and ordered from differences where sub-community 1 < 2 on the left to sub-community 1 > 2 on the right.

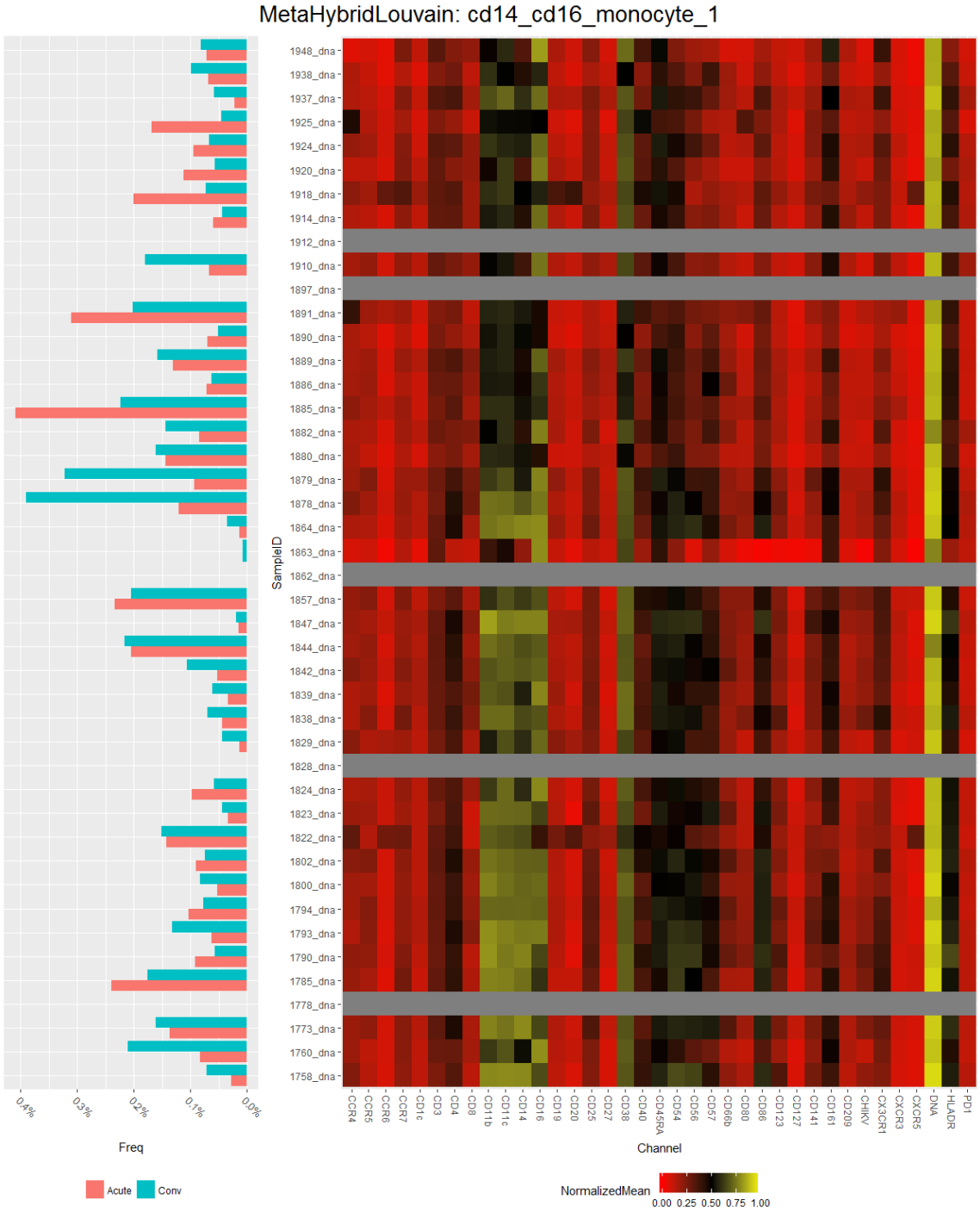


Figure B.4: **Summary of frequencies (per timepoint) and mean channel values for sub-community 1 of CD14$^+$CD16$^+$ monocytes**, across all samples. At left, frequencies in each sample split by timepoint; at right, mini-heatmaps of channel values for all events within this sub-community for each sample (plotted within each row). A gray row indicates this sub-community was not identified in this sample.

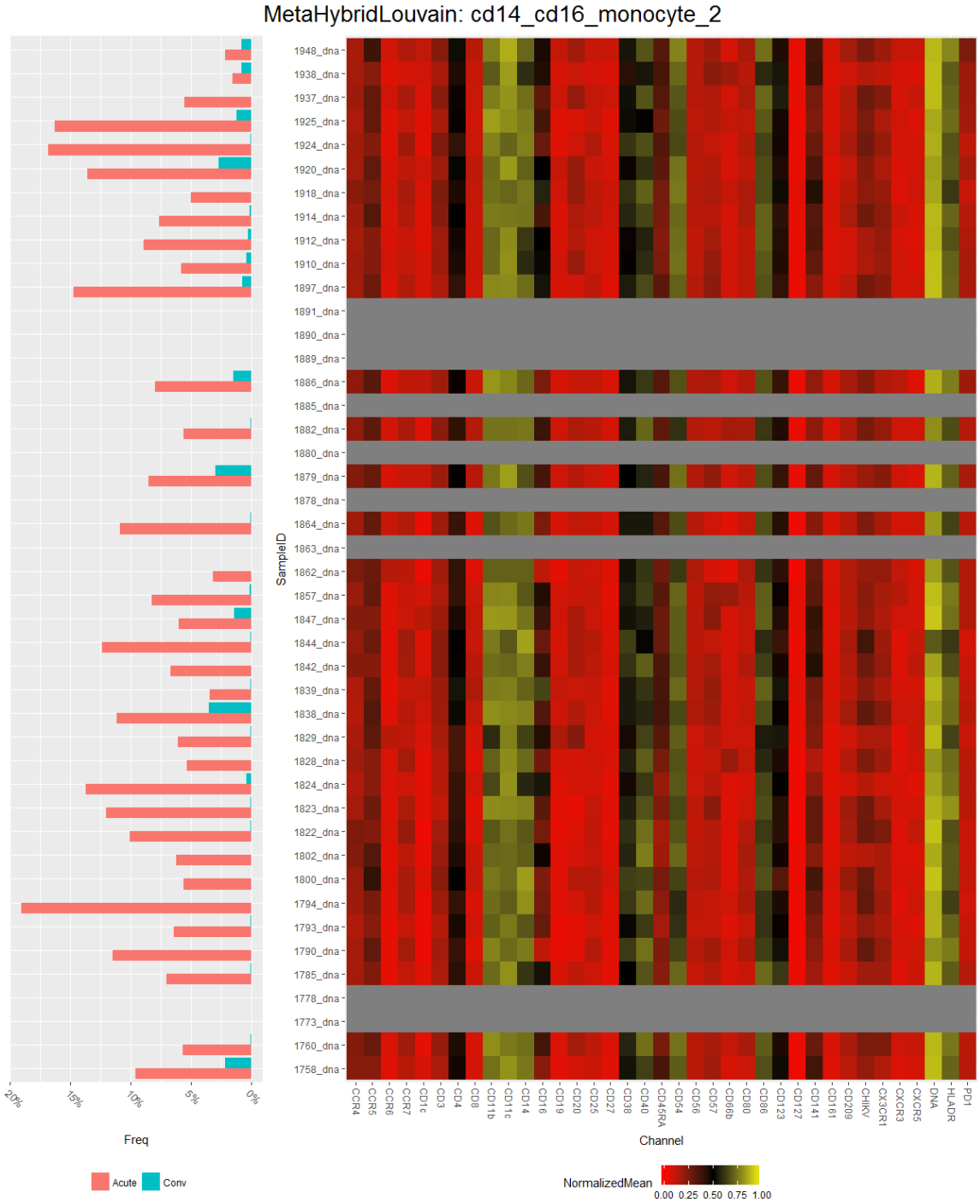


Figure B.5: **Summary of frequencies (per timepoint) and mean channel values for sub-community 2 of CD14$^+$CD16$^+$ monocytes**, across all samples. At left, frequencies in each sample split by timepoint; at right, mini-heatmaps of channel values for all events within this sub-community for each sample (plotted within each row). A gray row indicates this sub-community was not identified in this sample.

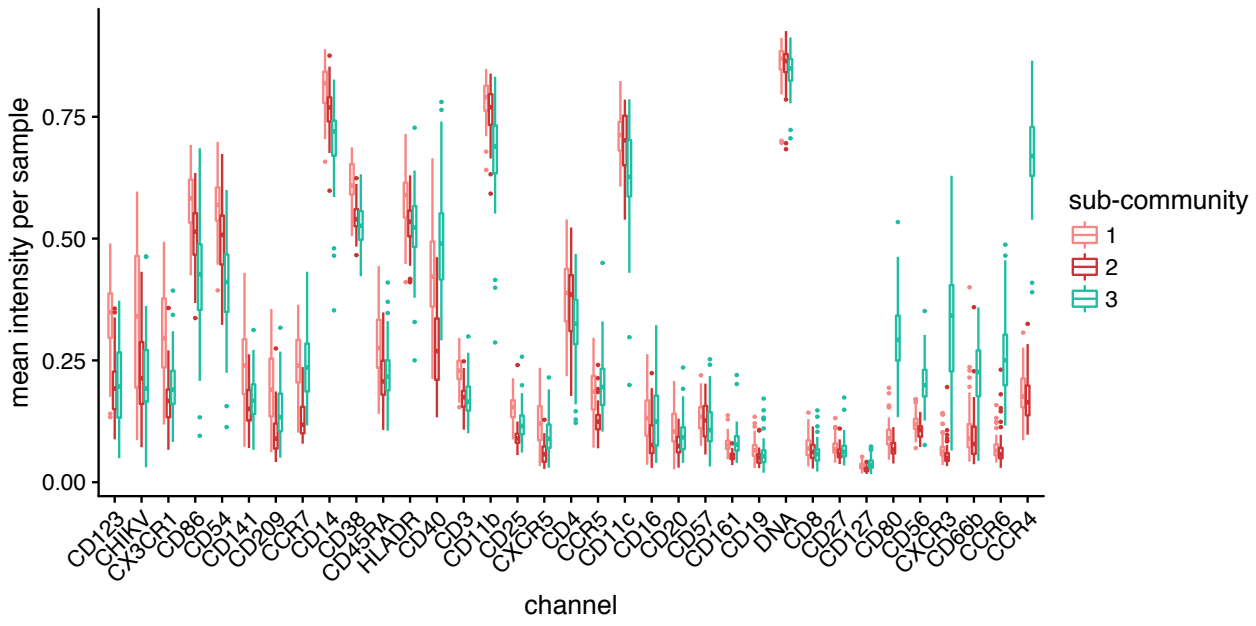


Figure B.6: **Differences in per-sample channel means between three CD14⁺ sub-communities identified by `MetaHybridLouvain`.** The X axis is filtered to only the channels with differences significant at FDR < 0.05 (Kruskal-Wallis test), and ordered from differences where sub-community 1 is greater than the other two sub-communities on the left to differences where sub-community 1 is lower than the other two sub-communities on the right.

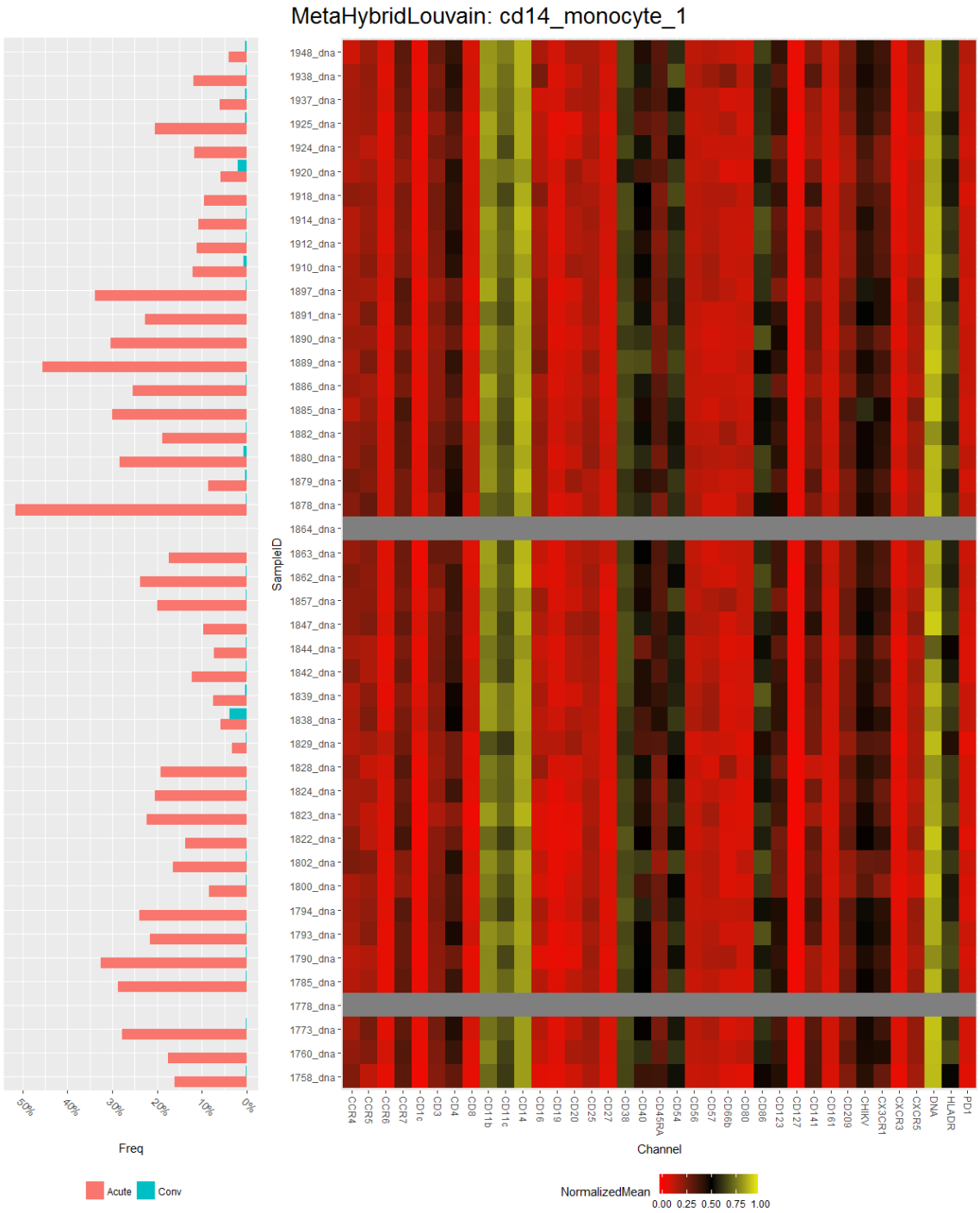


Figure B.7: **Summary of frequencies (per timepoint) and mean channel values for sub-community 1 of CD14$^+$ monocytes**, across all samples. At left, frequencies in each sample split by timepoint; at right, mini-heatmaps of channel values for all events within this sub-community for each sample (plotted within each row). A gray row indicates this sub-community was not identified in this sample.

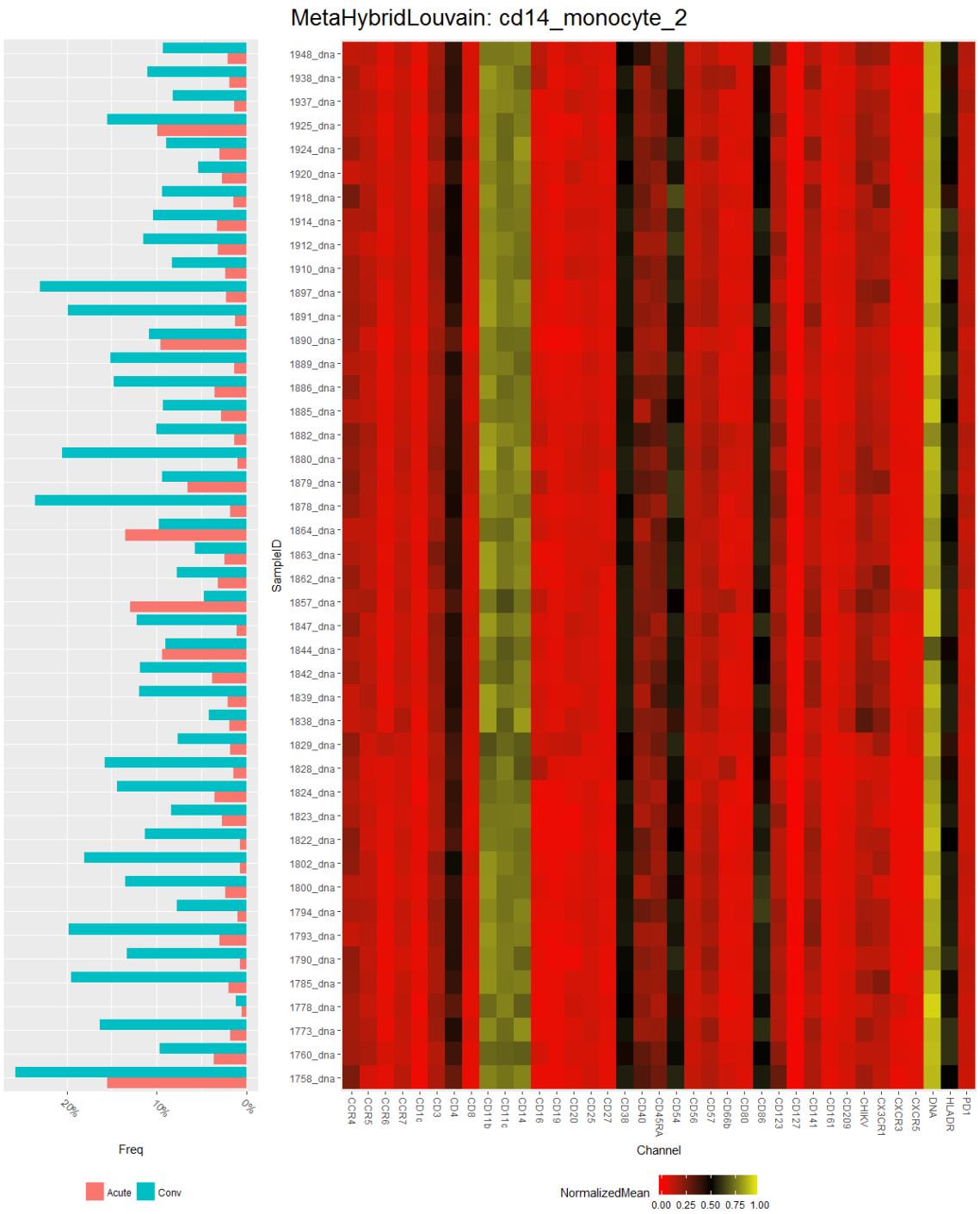


Figure B.8: **Summary of frequencies (per timepoint) and mean channel values for sub-community 2 of CD14$^+$ monocytes**, across all samples. At left, frequencies in each sample split by timepoint; at right, mini-heatmaps of channel values for all events within this sub-community for each sample (plotted within each row). A gray row indicates this sub-community was not identified in this sample.

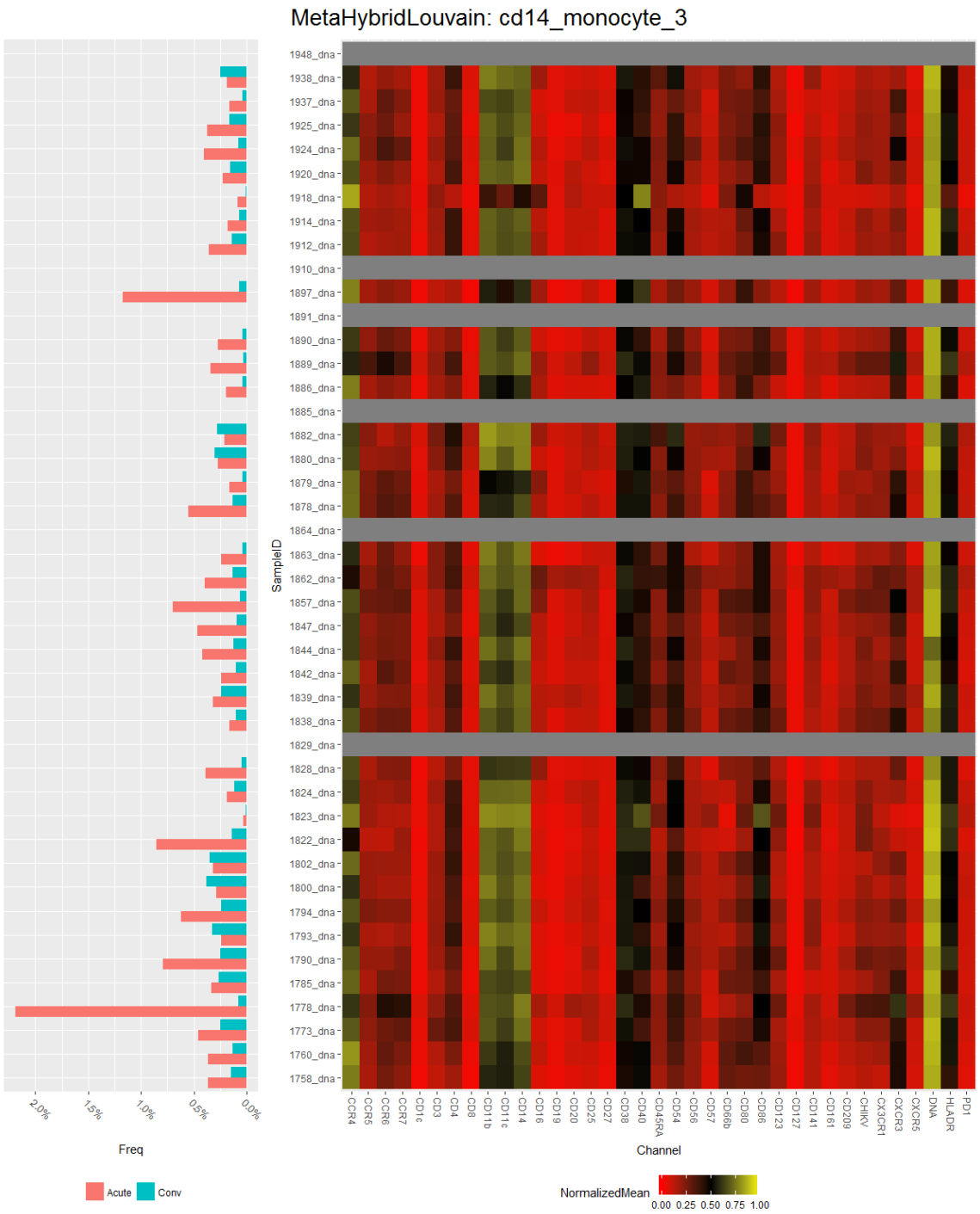


Figure B.9: **Summary of frequencies (per timepoint) and mean channel values for sub-community 3 of CD14$^+$ monocytes**, across all samples. At left, frequencies in each sample split by timepoint; at right, mini-heatmaps of channel values for all events within this sub-community for each sample (plotted within each row). A gray row indicates this sub-community was not identified in this sample.

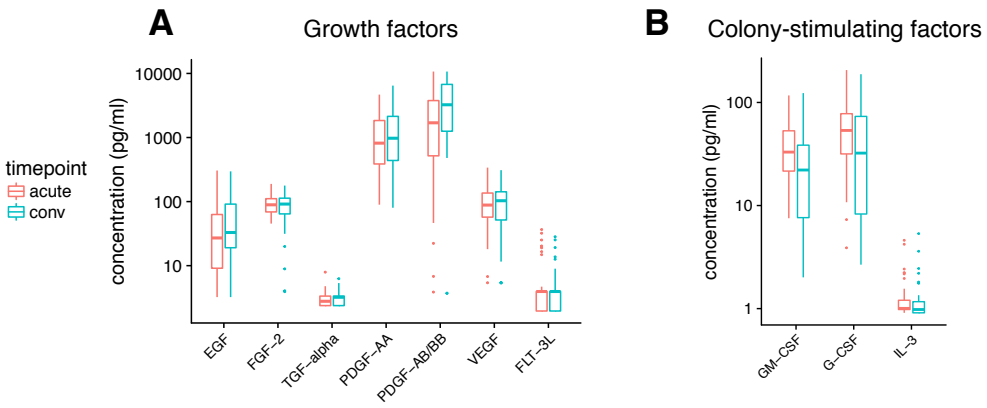


Figure B.10: **Differences in serum growth factor and colony-stimulating factor levels between the acute and convalescent timepoints**, as measured by multiplex ELISA (Luminex). None of the differences depicted here achieved statistical significance at FDR < 0.05 (Wilcoxon signed-rank test).

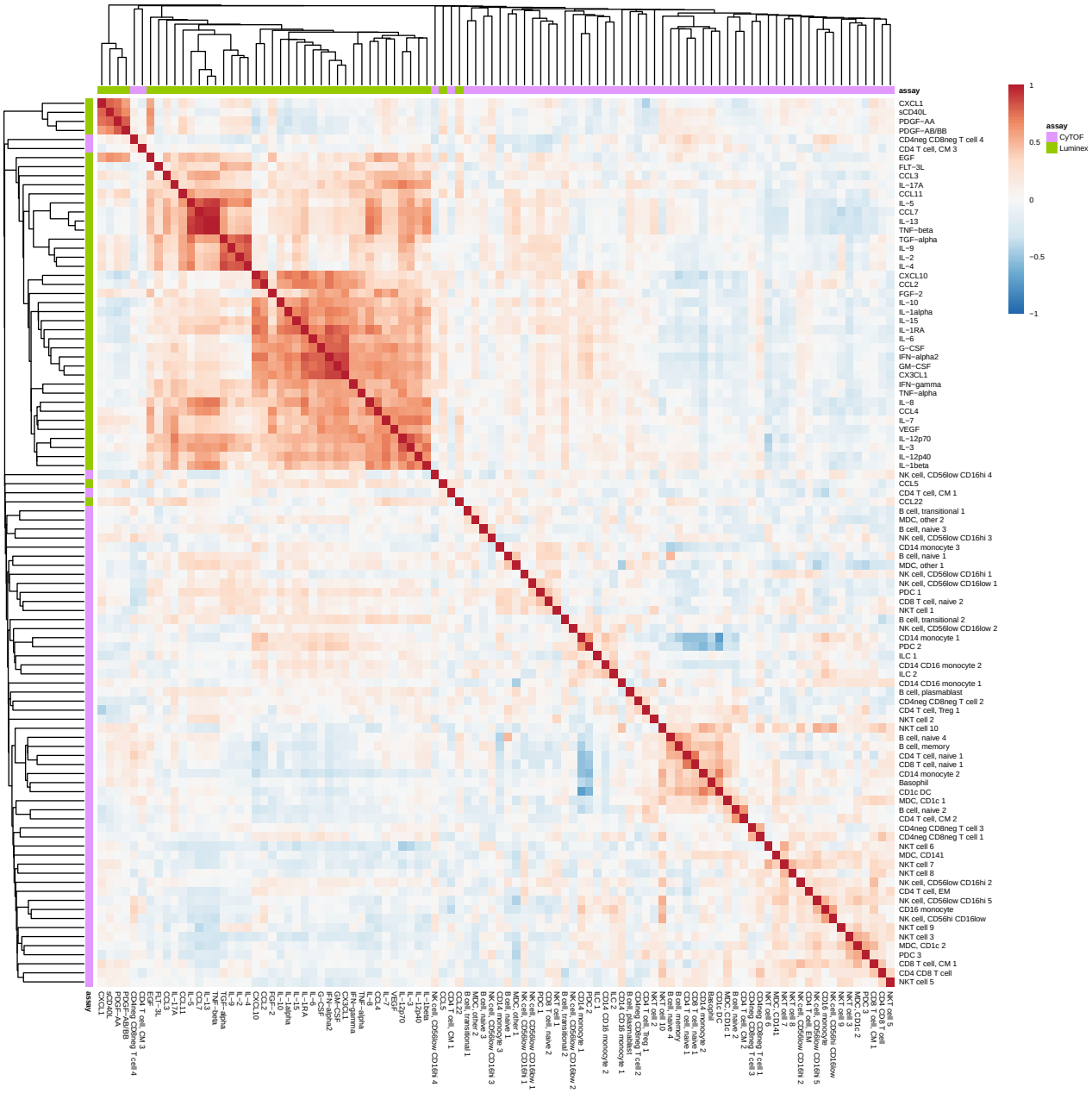


Figure B.11: **Clustered heatmap of Pearson correlations between log-scaled serum cytokine concentration (Luminex) and log-scaled cell subphenotype frequencies (CyTOF)** ***across the acute and convalescent timepoints.*** The source of each variable (CyTOF vs. Luminex) is depicted by the pink-green annotation bar.

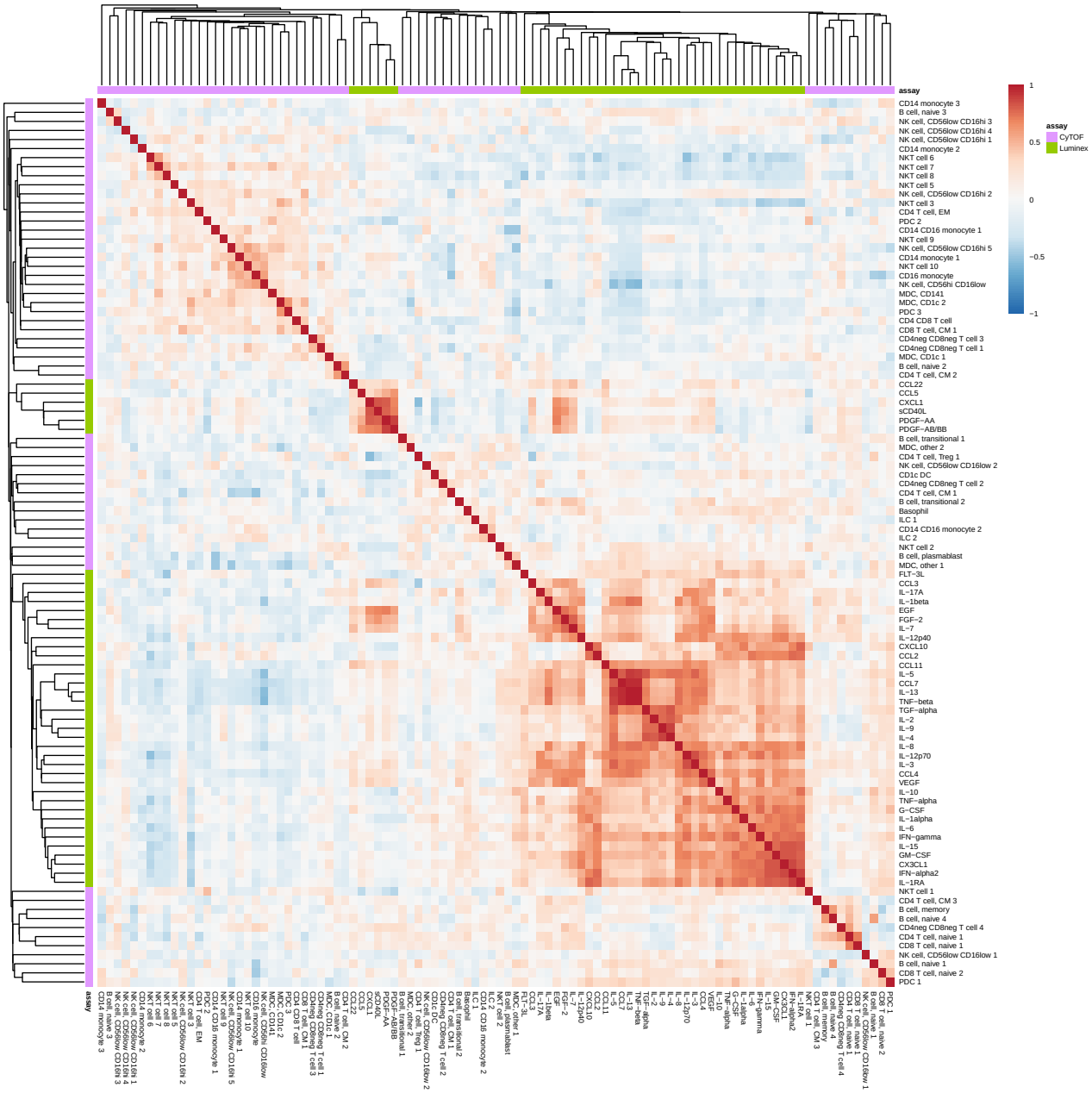


Figure B.12: **Clustered heatmap of Pearson correlations between log-scaled serum cytokine concentration (Luminex) and log-scaled cell subphenotype frequencies (CyTOF) *within the acute timepoint***. The source of each variable (CyTOF vs. Luminex) is depicted by the pink-green annotation bar.

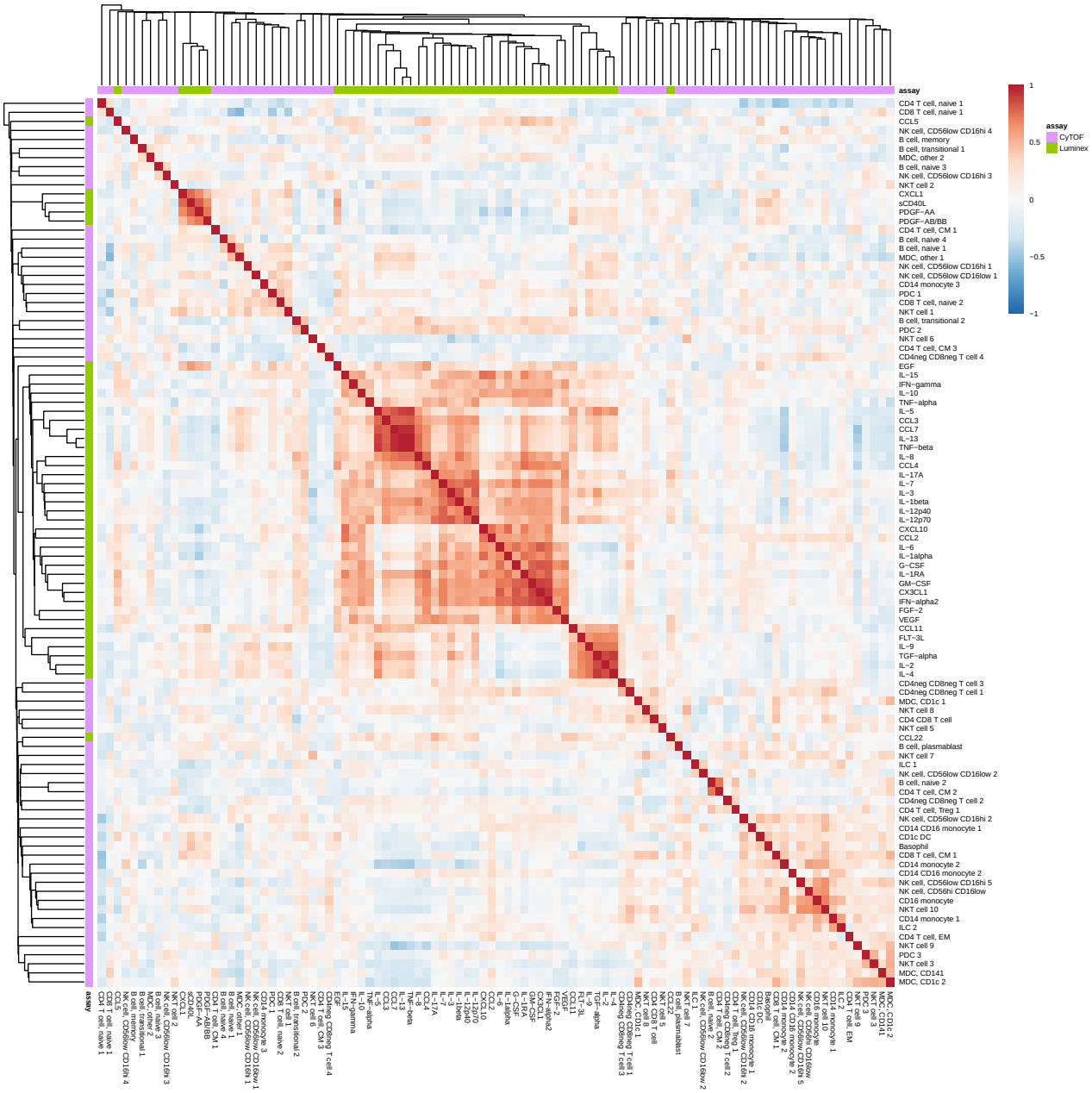


Figure B.13: **Clustered heatmap of Pearson correlations between log-scaled serum cytokine concentration (Luminex) and log-scaled cell subphenotype frequencies (CyTOF)** ***within the convalescent timepoint***. The source of each variable (CyTOF vs. Luminex) is depicted by the pink-green annotation bar.

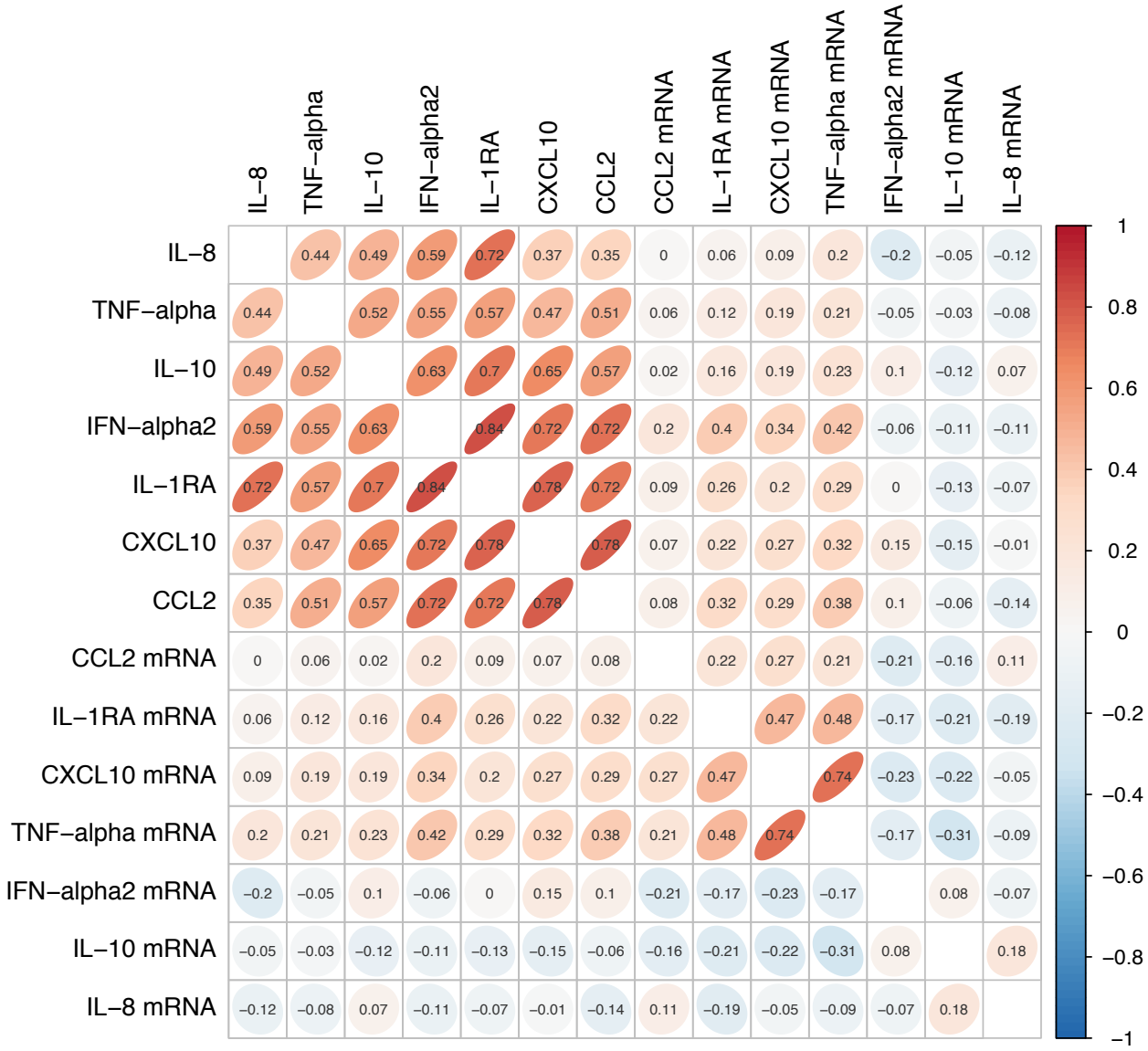


Figure B.14: **Pearson's correlations between serum cytokine concentrations that were significantly different between acute and convalescent timepoints and expression levels (in FPKM) for corresponding genes.** Both values were log scaled before calculating Pearson's *r*.

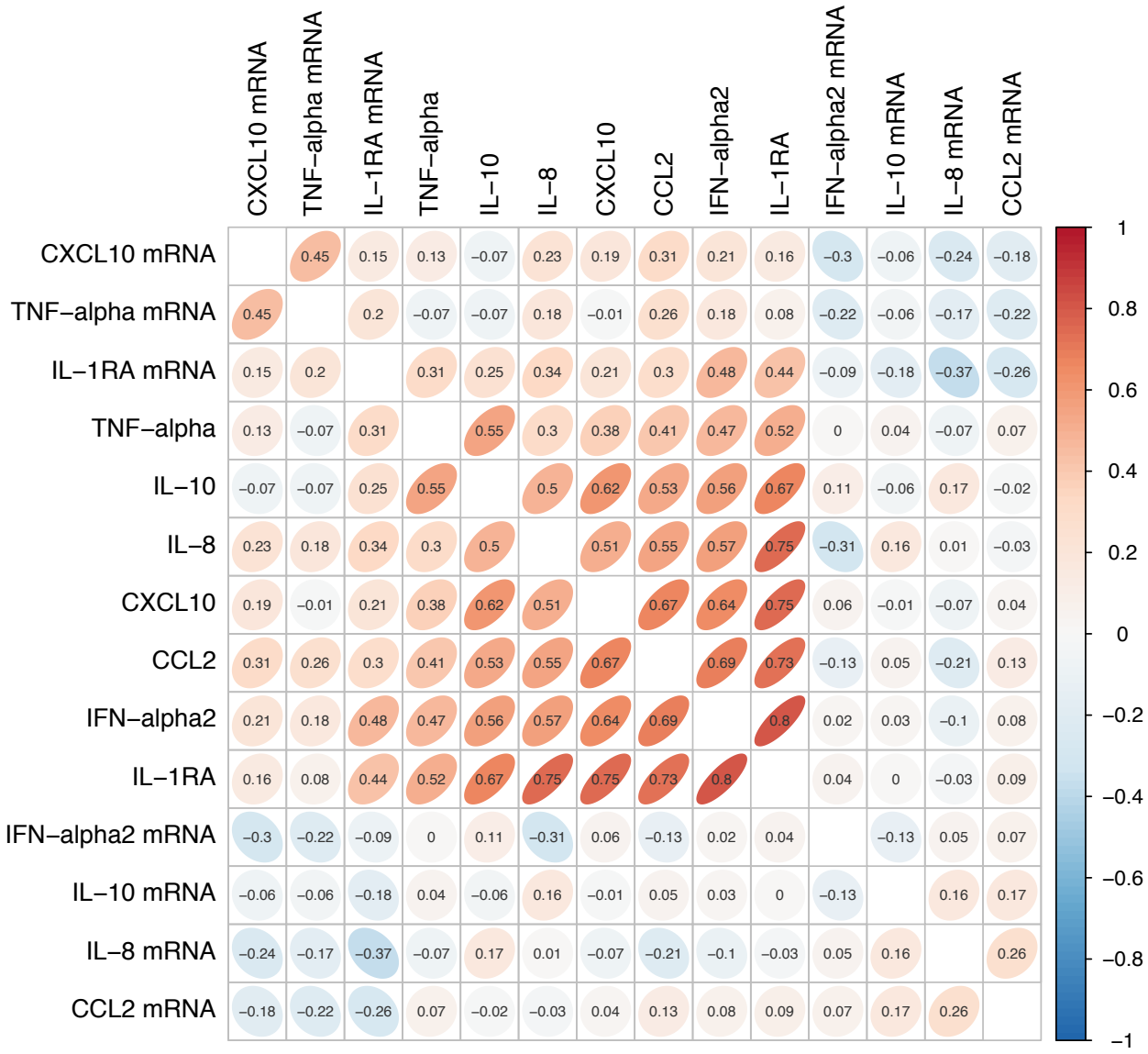


Figure B.15: **Pearson's correlations between serum cytokine concentrations that were significantly different between acute and convalescent timepoints and expression levels (in FPKM) for corresponding genes, normalizing all values at the acute timepoint against convalescent values.** Normalized values were log scaled before calculating Pearson's *r*.

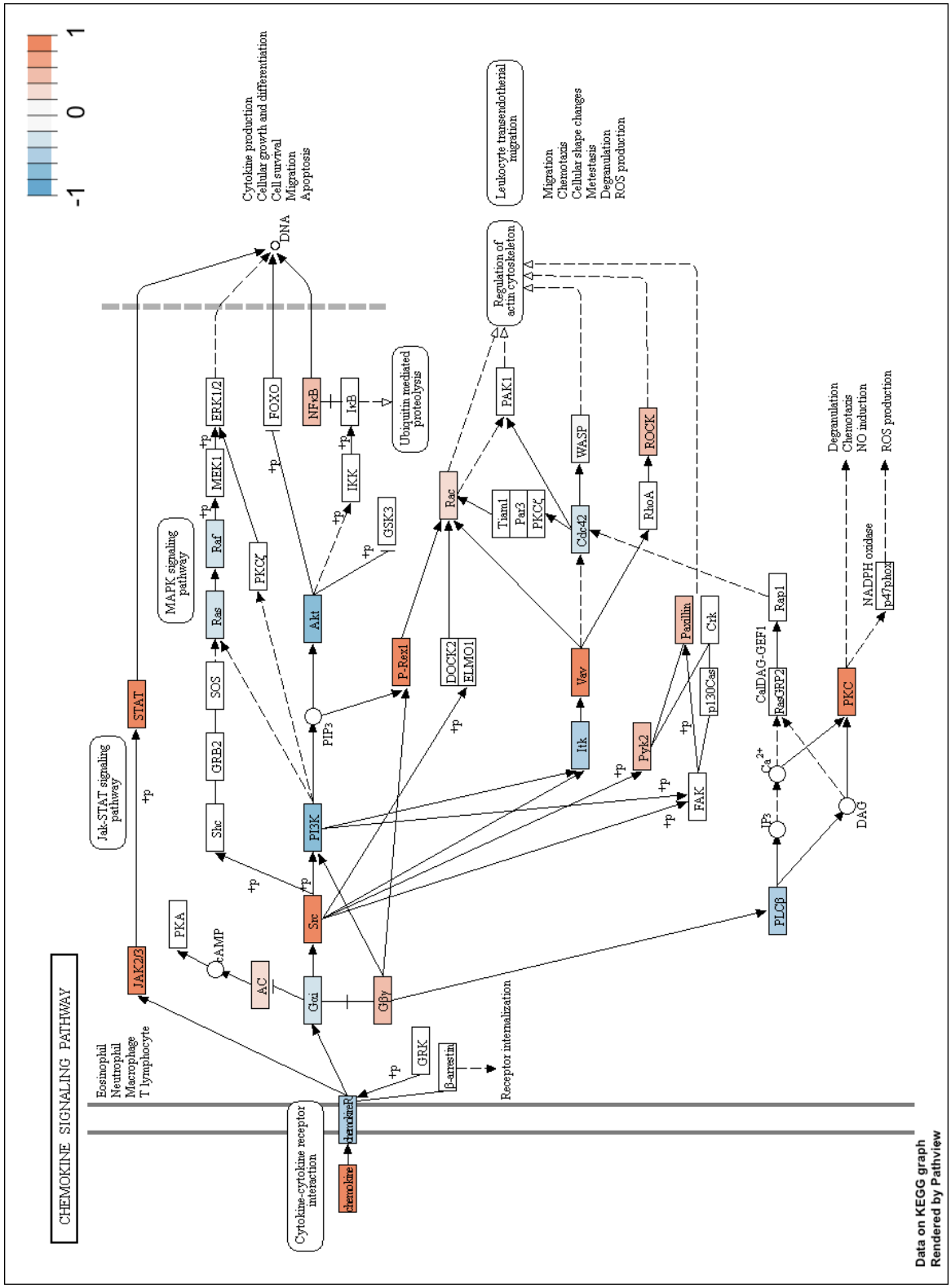


Figure B.16: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the chemokine signaling pathway**, using KEGG annotations (accession hsa04062). Positive values indicate upregulation during the acute phase of infection.

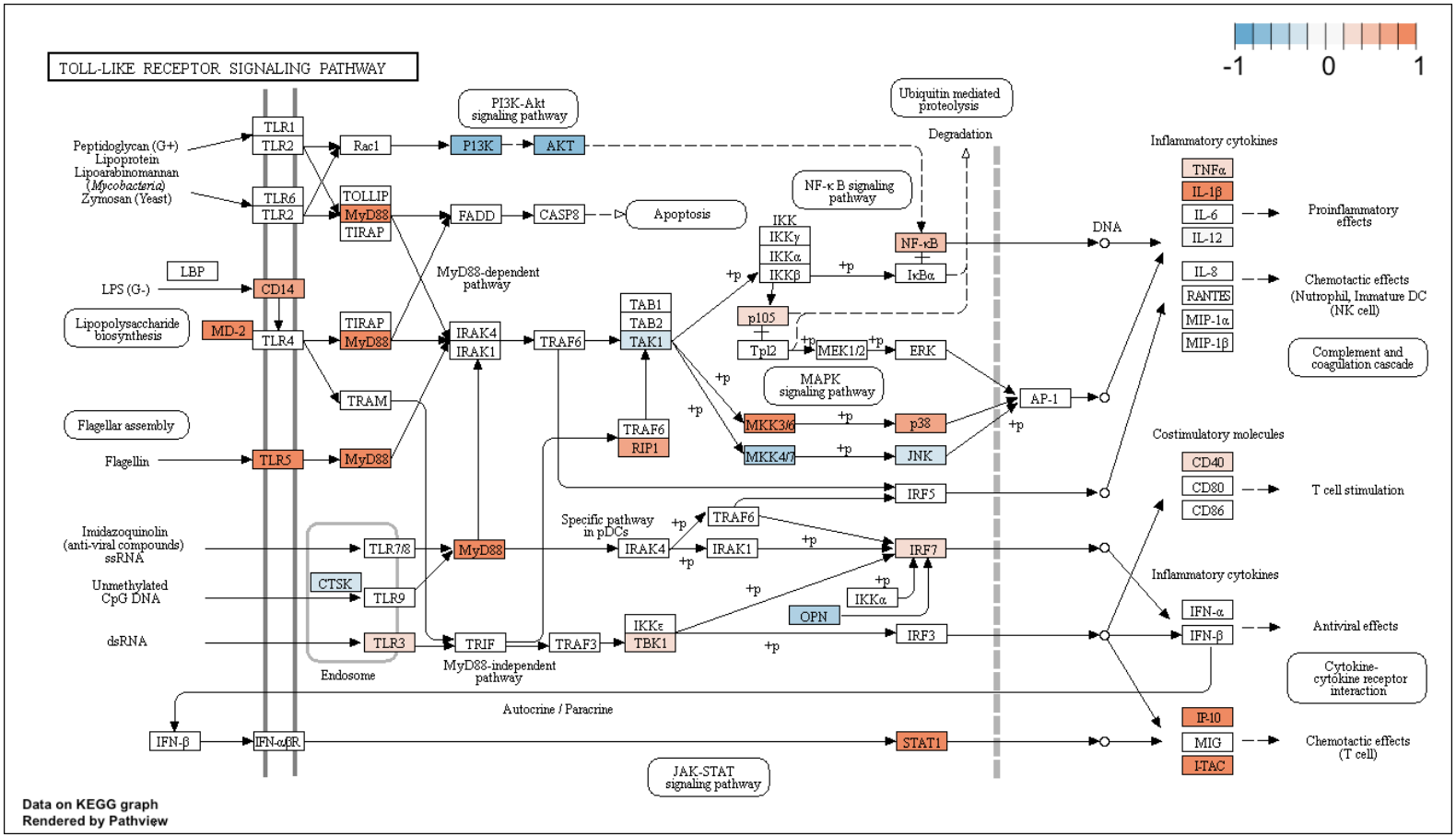


Figure B.17: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the toll-like receptor pathway**, using KEGG annotations (accession hsa04620). Positive values indicate upregulation during the acute phase of infection.

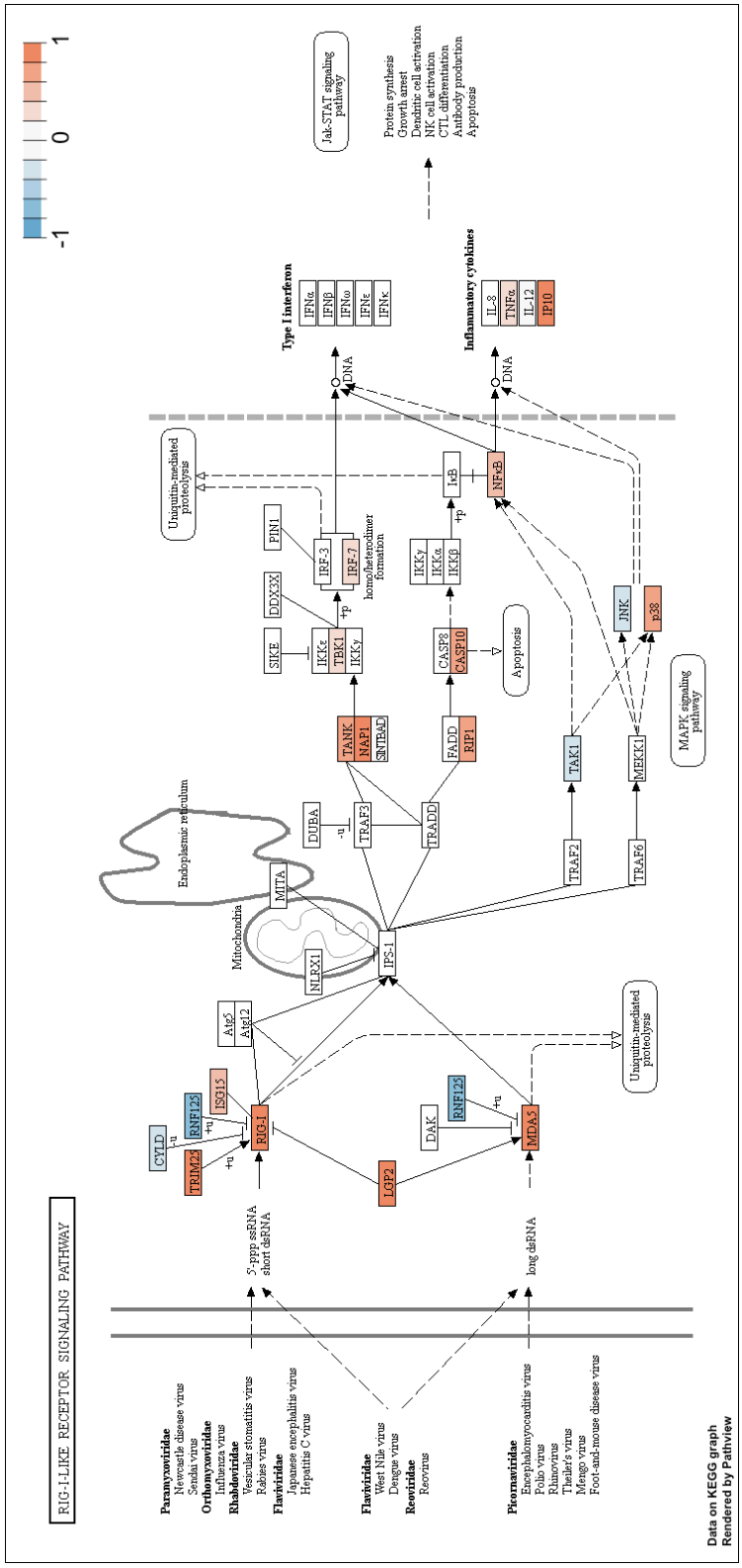


Figure B.18: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the RIG-I-like receptor signaling pathway**, using KEGG annotations (accession hsa04622). Positive values indicate upregulation during the acute phase of infection.

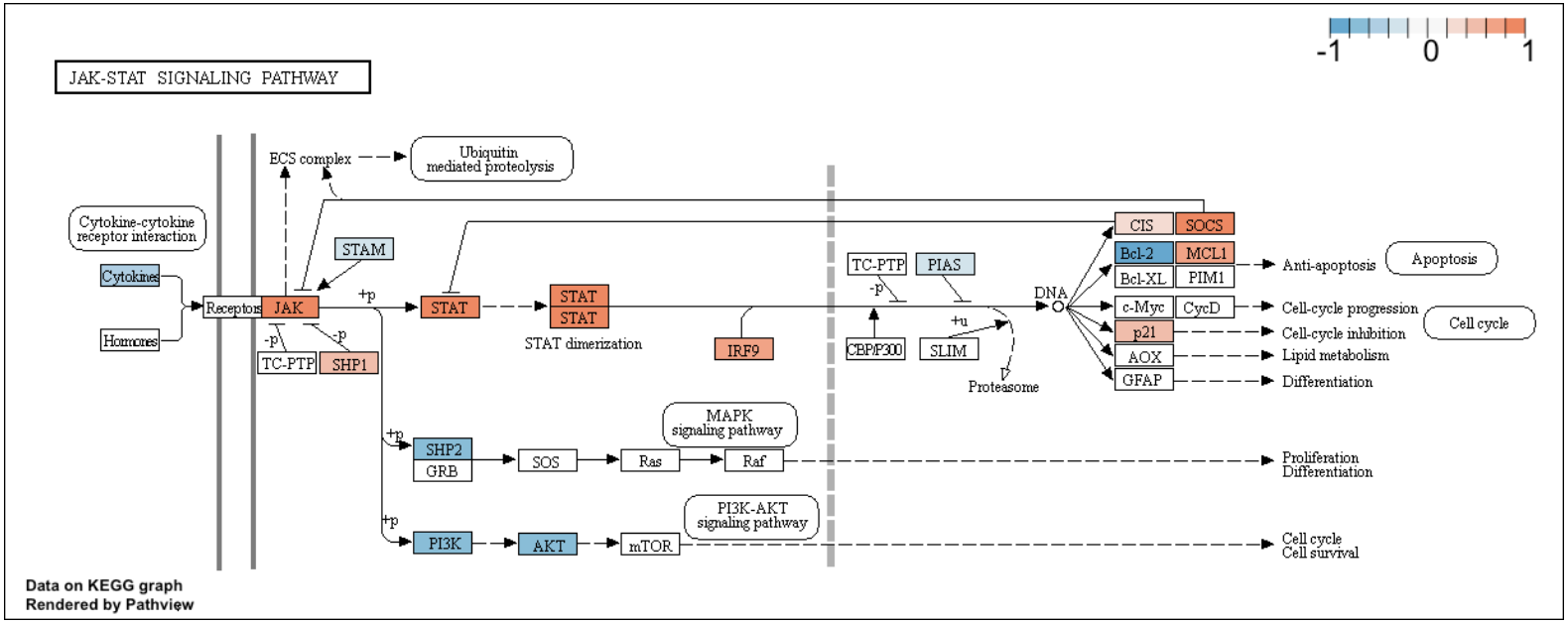


Figure B.19: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the JAK-STAT signaling pathway**, using KEGG annotations (accession hsa04630). Positive values indicate upregulation during the acute phase of infection.

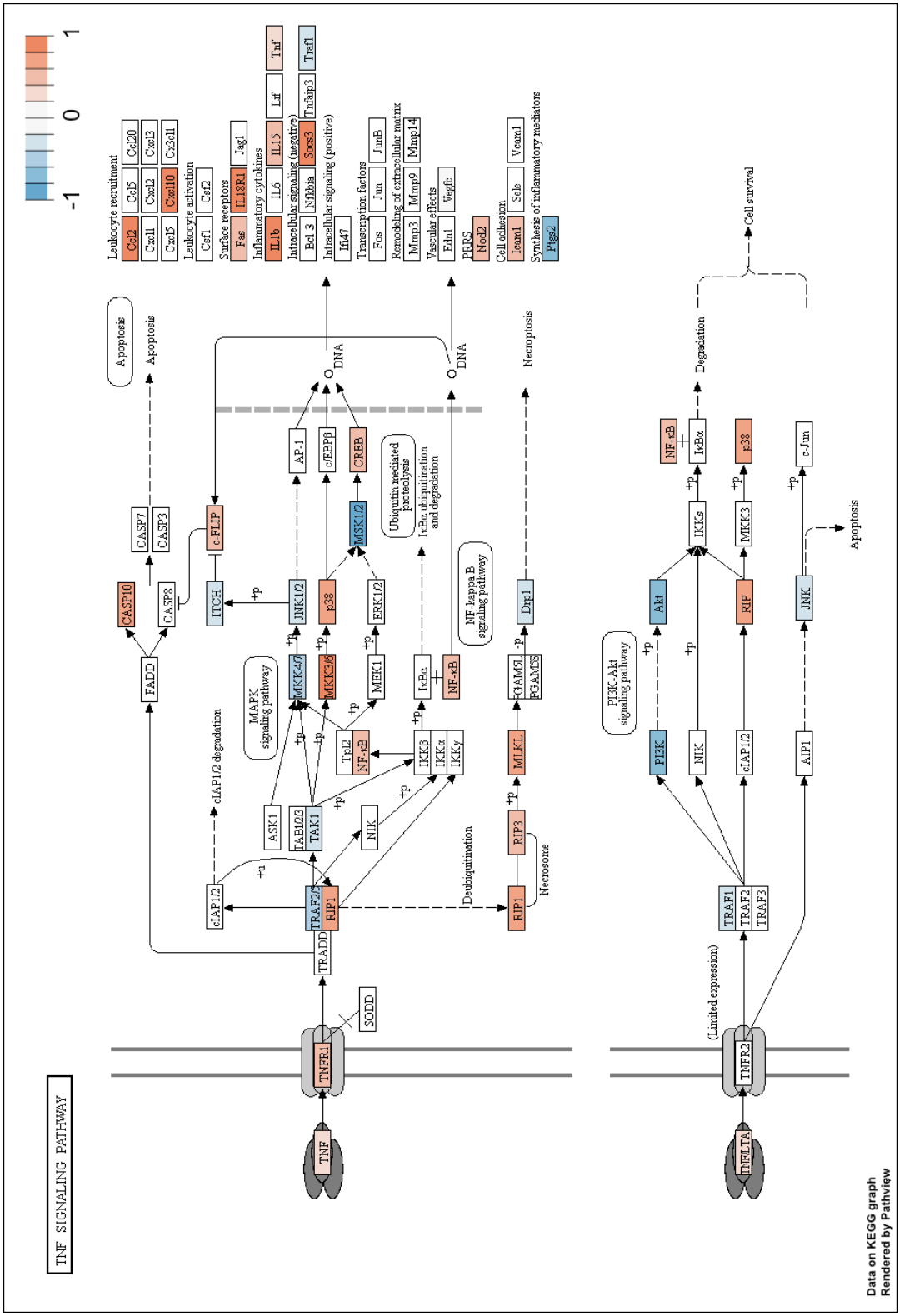


Figure B.20: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the TNF signaling pathway**, using KEGG annotations (accession hsa04668). Positive values indicate upregulation during the acute phase of infection.

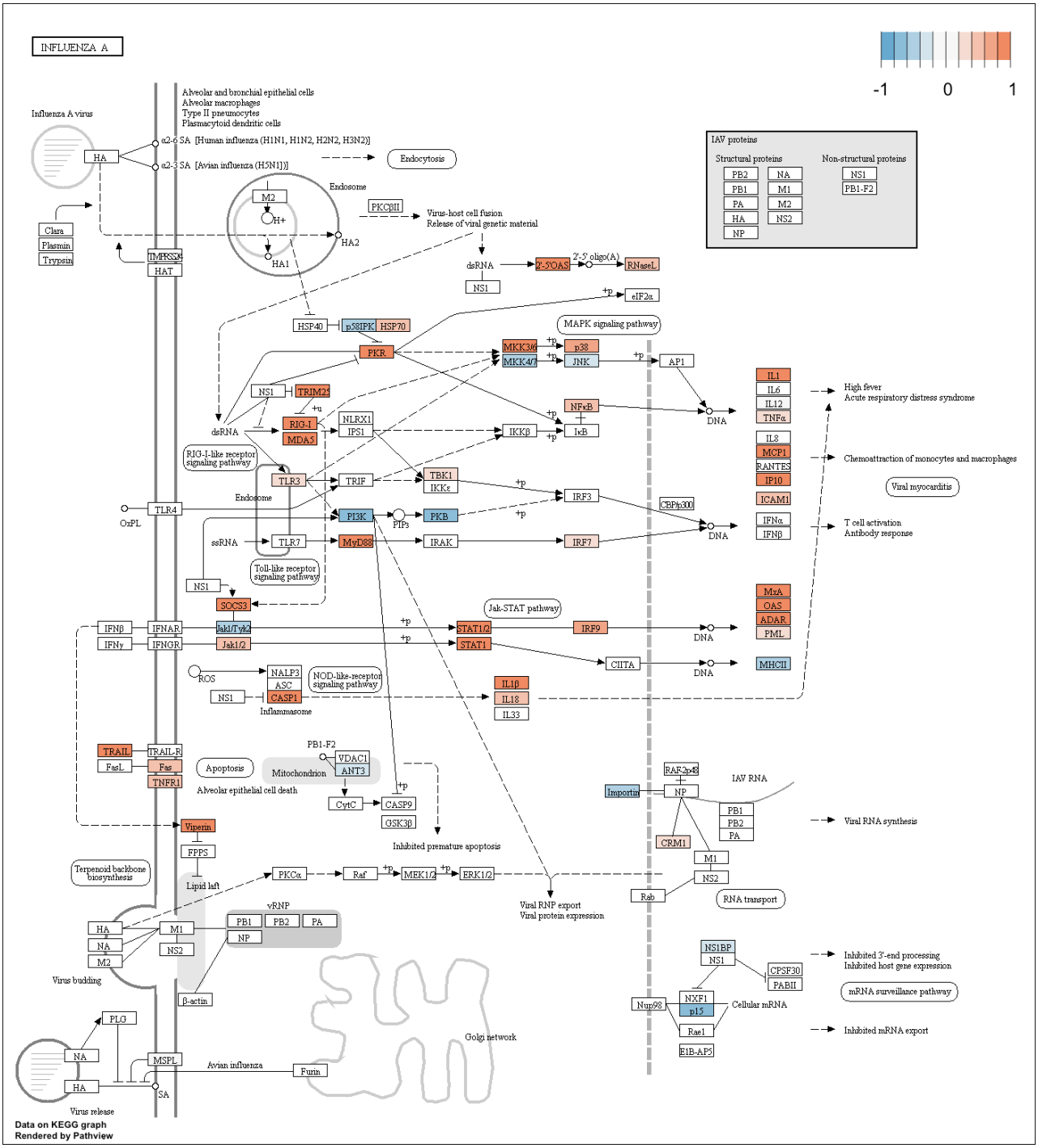


Figure B.21: **Pathview plot of $\log_2$ fold change in gene expression between acute and convalescent timepoints for the Influenza A pathway**, using KEGG annotations (accession hsa05164). Positive values indicate upregulation during the acute phase of infection.

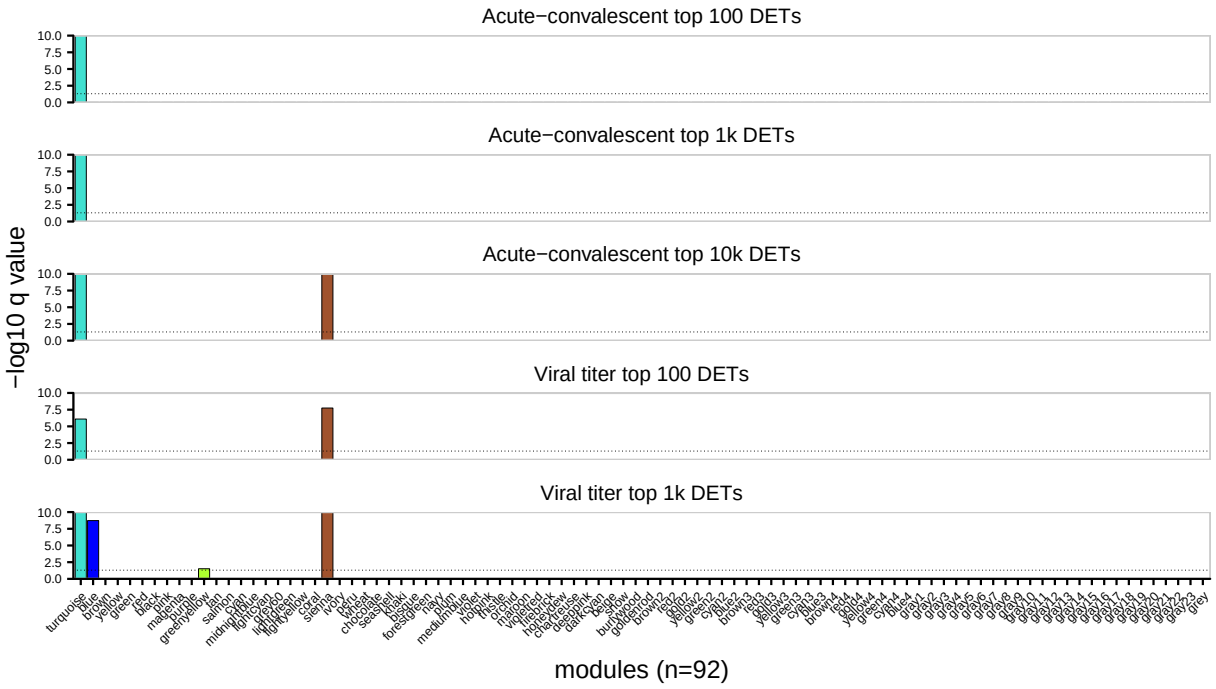


Figure B.22: ***q* values (Benjamini-Hochberg adjusted *P* values; Fisher's exact test) for enrichment analyses of five differentially expressed transcript signatures among the 92 coexpression modules**. Coexpression modules were created using whole genome coexpression network analysis; see Figure 6.25. The horizontal dashed line corresponds to a significance threshold of $q = 0.05$.

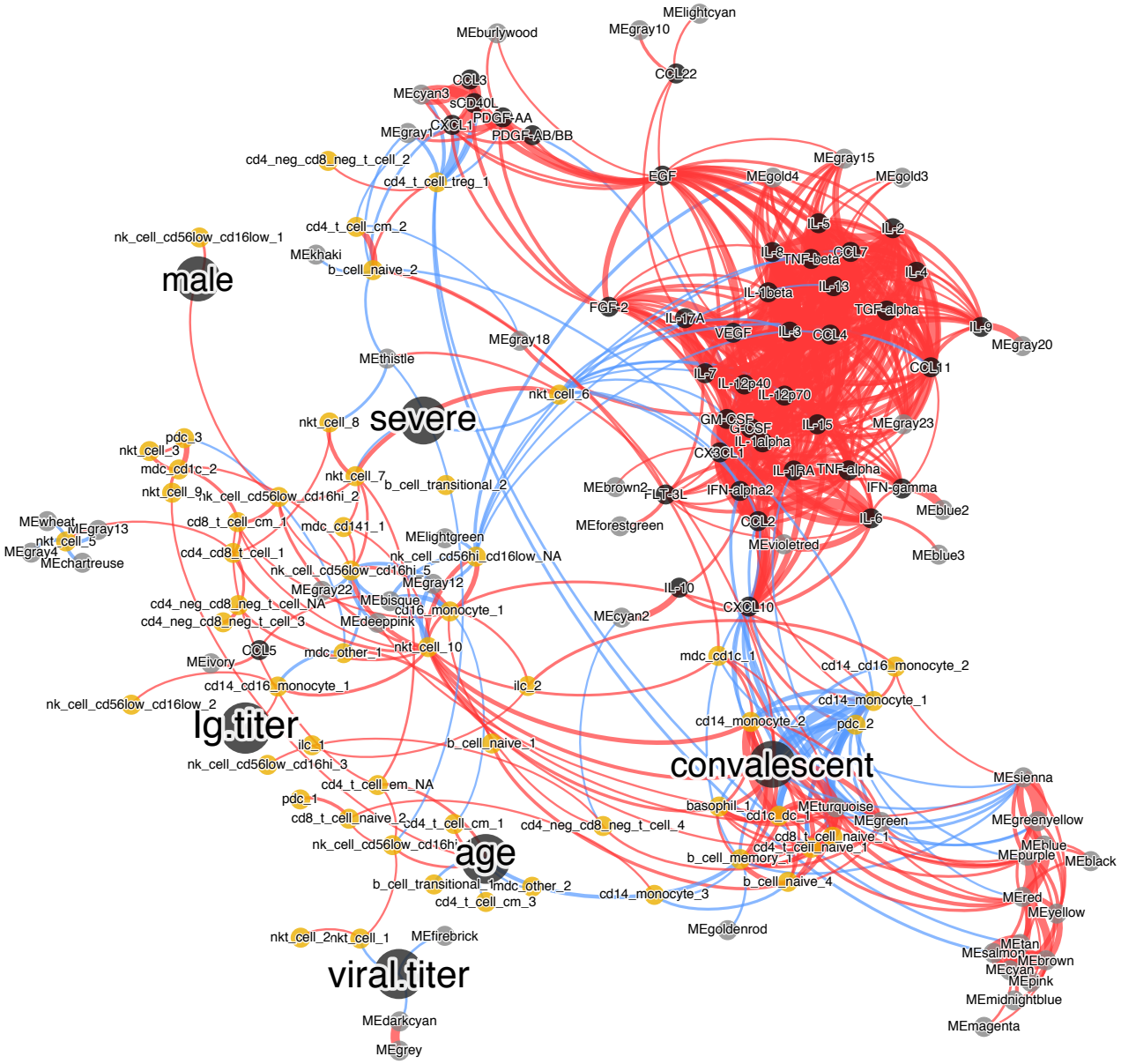


Figure B.23: **Weighted multiscale interaction network including Luminex data.** Edges represent all Pearson's correlations between subpopulation frequencies (gold nodes), serum cytokine concentrations (small black nodes), coexpression modules (gray nodes), and clinical variables (large black nodes), filtered to correlations significant at $P < 0.001$. Thickness is scaled to the magnitude of the correlation. The serum cytokines form a large interconnected component at top right that is not correlated with any of the clinical variables.

Bibliography

Aghaeepour N, Finak G, FlowCAP Consortium, et al. (2013). "Critical assessment of automated flow cytometry data analysis techniques." *Nat. Methods* 10.3, pp. 228–238. DOI: 10.1038/nmeth.2365.

Allen JD, Xie Y, Chen M, Girard L, and Xiao G (2012). "Comparing statistical methods for constructing large scale gene networks". *PLoS One* 7.1, pp. 17–19. DOI: 10.1371/journal.pone.0029348.

Allignol A, Schumacher M, and Beyersmann J (2011). "Empirical Transition Matrix of Multi-State Models: The etm Package". *J. Stat. Softw.* 38.4, pp. 1–15. DOI: 10.18637/jss.v038.i04.

Alonso A and Martínez JL (1997). "Multiple antibiotic resistance in Stenotrophomonas maltophilia." *Antimicrob. Agents Chemother.* 41.5, pp. 1140–1142.

Alonso A and Martínez JL (2001). "Expression of multidrug efflux pump smeDEF by clinical isolates of Stenotrophomonas maltophilia". *Antimicrob. Agents Chemother.* 45.6, pp. 1879–1881. DOI: 10.1128/AAC.45.6.1879-1881.2001.

Alonso A, Morales G, Escalante R, et al. (2004). "Overexpression of the multidrug efflux pump SmeDEF impairs Stenotrophomonas maltophilia physiology". *J. Antimicrob. Chemother.* 53.3, pp. 432–434. DOI: 10.1093/jac/dkh074.

Altman DR, Sebra R, Hand J, et al. (2014). "Transmission of Methicillin-Resistant Staphylococcus aureus via Deceased Donor Liver Transplantation Confirmed by Whole Genome Sequencing." *Am. J. Transplant* 14.11, pp. 2640–4. DOI: 10.1111/ajt.12897.

Amir EaD, Davis KL, Tadmor MD, et al. (2013). "viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia." *Nat. Biotechnol.* 31.6, pp. 545–52. DOI: 10.1038/nbt.2594.

Anders S and Huber W (2010). "Differential expression analysis for sequence count data." *Genome Biol.* 11.10, R106. DOI: 10.1186/gb-2010-11-10-r106.

Anders S, Pyl PT, and Huber W (2015). "HTSeq-A Python framework to work with high-throughput sequencing data". *Bioinformatics* 31.2, pp. 166–169. DOI: 10.1093/bioinformatics/btu638.

Anders S, Reyes A, and Huber W (2012). "Detecting differential usage of exons from RNA-seq data." *Genome Res.* 22.10, pp. 2008–17. DOI: 10.1101/gr.133744.111.

Anderson DJ, Harris SR, Godofsky E, et al. (2014). "Whole Genome Sequencing of a Methicillin-Resistant Staphylococcus aureus Pseudo-Outbreak in a Professional Football Team." *Open Forum Infect. Dis.* 1.3, ofu096. DOI: 10.1093/ofid/ofu096.

Angiuoli SV and Salzberg SL (2011). "Mugsy: fast multiple alignment of closely related whole genomes." *Bioinformatics* 27.3, pp. 334–42. DOI: 10.1093/bioinformatics/btq665.

Angiuoli SV, Matalka M, Gussman A, et al. (2011). "CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing." *BMC Bioinformatics* 12.1, p. 356. DOI: 10.1186/1471-2105-12-356.

Appleby LJ, Nausch N, Midzi N, et al. (2013). "Sources of heterogeneity in human monocyte subsets". *Immunol. Lett.* 152.1, pp. 32–41. DOI: 10.1016/j.imlet.2013.03.004.

Arazi A, Pendergraft WF, Ribeiro RM, Perelson AS, and Hacohen N (2013). "Human systems immunology: Hypothesis-based modeling and unbiased data-driven approaches". *Semin. Immunol.* 25.3, pp. 193–200. DOI: 10.1016/j.smim.2012.11.003.

Assunção-Miranda I, Cruz-Oliveira C, and Da Poian AT (2013). "Molecular mechanisms involved in the pathogenesis of alphavirus-induced arthritis." *Biomed Res. Int.* 2013, p. 973516. DOI: 10.1155/2013/973516.

Austin PC (2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behav. Res.* 46.3, pp. 399–424. DOI: 10.1080/00273171.2011.568786.

Azarian T, Cook RL, Johnson JA, et al. (2015). "Whole-genome sequencing for outbreak investigations of methicillin-resistant Staphylococcus aureus in the neonatal intensive care unit: time for routine practice?" *Infect. Control Hosp. Epidemiol.* 36.7, pp. 777–85. DOI: 10.1017/ice.2015.73.

Aziz RK, Bartels D, Best AA, et al. (2008). "The RAST Server: rapid annotations using subsystems technology." *BMC Genomics* 9.1, p. 75. DOI: 10.1186/1471-2164-9-75.

Bagdasarian N, Rao K, and Malani PN (2015). "Diagnosis and Treatment of Clostridium difficile in Adults". *JAMA* 313.4, p. 398. DOI: 10.1001/jama.2014.17103.

Bankevich A, Nurk S, Antipov D, et al. (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing". *J. Comput. Biol.* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021.

Bashir A, Klammer AA, Robins WP, et al. (2012). "A hybrid approach for the automated finishing of bacterial genomes". *Nat. Biotechnol.* 30.7, pp. 701–7. DOI: 10.1038/nbt.2288.

Bashir A, Attie O, Sullivan M, et al. (2017). "Genomic confirmation of vancomycin-resistant Enterococcus transmission from deceased donor to liver transplant recipient." *PLoS One* 12.3, e0170449. DOI: 10.1371/journal.pone.0170449.

Bastian M, Heymann S, and Jacomy M (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". *Third Int. AAAI Conf. Weblogs Soc. Media*, pp. 361–362. DOI: 10.1136/qshc.2004.010033.

Beaulaurier J, Zhang XS, Zhu S, et al. (2015). "Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes". *Nat. Commun.* 6, p. 7438. DOI: 10.1038/ncomms8438.

Benjamini Y and Yekutieli D (2001). "The control of the false discovery rate in multiple testing under depencency". *Ann. Stat.* 29.4, pp. 1165–1188.

Best EL, Parnell P, Thirkell G, et al. (2014). "Effectiveness of deep cleaning followed by hydrogen peroxide decontamination during high Clostridium difficile infection incidence". *J. Hosp. Infect.* 87.1, pp. 25–33. DOI: 10.1016/j.jhin.2014.02.005.

Boehme CC, Nabeta P, Hillemann D, et al. (2010). "Rapid molecular detection of tuberculosis and rifampin resistance." *N. Engl. J. Med.* 363.11, pp. 1005–15. DOI: 10.1056/NEJMoa0907847.

Bradley P, Gordon NC, Walker TM, et al. (2015). "Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis". *Nat. Commun.* 6, p. 10063. DOI: 10.1038/ncomms10063.

Bray NL, Pimentel H, Melsted P, and Pachter L (2016). "Near-optimal probabilistic RNA-seq quantification". *Nat. Biotechnol.* 34.5, pp. 525–527. DOI: 10.1038/nbt.3519.

Brooke JS (2012). "Stenotrophomonas maltophilia: an emerging global opportunistic pathogen." *Clin. Microbiol. Rev.* 25.1, pp. 2–41. DOI: 10.1128/CMR.00019-11.

Brownstein JS, Freifeld CC, Reis BY, and Mandl KD (2008). "Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project." *PLoS Med.* 5.7, e151. DOI: 10.1371/journal.pmed.0050151.

Brynildsrud O, Bohlin J, Scheffer L, and Eldholm V (2016). "Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary". *Genome Biol.* 17.1, p. 238. DOI: 10.1186/s13059-016-1108-8.

Buechner JS, Constantine H, and Gjelsvik A (2004). "John Snow and the Broad Street pump: 150 years of epidemiology." *Med. Health. R. I.* 87.10, pp. 314–5.

Buels R, Yao E, Diesh CM, et al. (2016). "JBrowse: a dynamic web platform for genome visualization and analysis". *Genome Biol.* 17.1, pp. 1–12. DOI: 10.1186/s13059-016-0924-1.

Burt FJ, Chen W, Miner JJ, et al. (2017). "Chikungunya virus : an update on the biology and pathogenesis of this emerging pathogen". *Lancet Infect. Dis.* 3099.16, pp. 1–11. DOI: 10.1016/S1473-3099(16)30385-1.

Casali N, Broda A, Harris SR, et al. (2016). "Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study". *PLOS Med.* 13.10, e1002137. DOI: 10.1371/journal.pmed.1002137.

Casas V, Miyake J, Balsley H, et al. (2006). "Widespread occurrence of phage-encoded exotoxin genes in terrestrial and aquatic environments in Southern California." *FEMS Microbiol. Lett.* 261.1, pp. 141–9. DOI: 10.1111/j.1574-6968.2006.00345.x.

Chaaitanya IK, Muruganandam N, Sundaram SG, et al. (2011). "Role of proinflammatory cytokines and chemokines in chronic arthropathy in CHIKV infection." *Viral Immunol.* 24.4, pp. 265–71. DOI: 10.1089/vim.2010.0123.

Chaisson MJ and Tesler G (2012). "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory." *BMC Bioinformatics* 13, p. 238. DOI: 10.1186/1471-2105-13-238.

Charron L, Doctrinal A, Choileain SN, and Astier AL (2015). "Monocyte : T-cell interaction regulates human T-cell activation through a CD28 / CD46 crosstalk". *Immunol. Cell Biol.* 93.9, pp. 796–803. DOI: 10.1038/icb.2015.42.

Check Hayden E (2014). "Data from pocket-sized genome sequencer unveiled". *Nat. News* 2014.Feb14. DOI: 10.1038/nature.2014.14724.

Chen EY, Tan CM, Kou Y, et al. (2013). "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool." *BMC Bioinformatics* 14.1, p. 128. DOI: 10.1186/1471-2105-14-128.

Chen LM, Davis CT, Zhou H, Cox NJ, and Donis RO (2008). "Genetic Compatibility and Virulence of Reassortants Derived from Contemporary Avian H5N1 and Human H3N2 Influenza A Viruses". *PLoS Pathog.* 4.5, e1000072.

Chen L, Mathema B, Chavda KD, et al. (2014). "Carbapenemase-producing Klebsiella pneumoniae: molecular and genetic decoding." *Trends Microbiol.* Pp. 1–11. DOI: 10.1016/j.tim.2014.09.003.

Chen PE and Shapiro BJ (2015). "The advent of genome-wide association studies for bacteria". *Curr. Opin. Microbiol.* 25, pp. 17–24. DOI: 10.1016/j.mib.2015.03.002.

Chin CS, Sorenson J, Harris JB, et al. (2011). "The Origin of the Haitian Cholera Outbreak Strain". *N. Engl. J. Med.* 364.1, pp. 33–42. DOI: 10.1056/NEJMoa1012928.

Chin CS, Alexander DH, Marks P, et al. (2013). "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data." *Nat. Methods* 10.6, pp. 563–9. DOI: 10.1038/nmeth.2474.

Cho SY, Kang CI, Kim J, et al. (2014). "Can levofloxacin be a useful alternative to trimethoprim-sulfamethoxazole for treating Stenotrophomonas maltophilia bacteremia?" *Antimicrob. Agents Chemother.* 58.1, pp. 581–583. DOI: 10.1128/AAC.01682-13.

Chow A, Her Z, Ong EKS, et al. (2011). "Persistent arthralgia induced by Chikungunya virus infection is associated with interleukin-6 and granulocyte macrophage colony-stimulating factor." *J. Infect. Dis.* 203.2, pp. 149–157. DOI: 10.1093/infdis/jiq042.

Clark TA, Murray IA, Morgan RD, et al. (2012). "Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing". *Nucleic Acids Res.* 40.4. DOI: 10.1093/nar/gkr1146.

Clinical and Laboratory Standards Institute (2015). *Performance standards for antimicrobial susceptibility testing; twenty-fifth informational supplement M100-S25*. Wayne, PA: Clinical and Laboratory Standards Institute.

Cohen SH, Gerding DN, Johnson S, et al. (2010). "Clinical practice guidelines for Clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of America (SHEA) and the infectious diseases society of America (IDSA)." *Infect. Control Hosp. Epidemiol.* 31.5, pp. 431–55. DOI: 10.1086/651706.

Comas I, Borrell S, Roetzer A, et al. (2011). "Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes". *Nat. Genet.* 44.1, pp. 106–110. DOI: 10.1038/ng.1038.

Committee on the Learning Healthcare System in America and Institute of Medicine (2014). *Best care at lower cost: the path to continuously learning health care in America*. Ed. by M Smith, R Saunders, L Stuckhardt, and JM McGinnis. Vol. 51. 06. Washington, DC: The National Academies Press. DOI: 10.5860/CHOICE.51-3277.

Cooper Z, Craig S, Gaynor M, and Van Reenen J (2015). "The Price Ain't Right? Hospital Prices and Health Spending on the Privately Insured". *NBER Work. Pap.* NBER Working Paper, p. 21815. DOI: 10.3386/w21815.

Costa V, Aprile M, Esposito R, and Ciccodicola A (2012). "RNA-Seq and human complex diseases: recent accomplishments and future perspectives". *Eur. J. Hum. Genet.* 21.10, pp. 134–142. DOI: 10.1038/ejhg.2012.129.

Couderc T and Lecuit M (2015). "Chikungunya virus pathogenesis: From bedside to bench". *Antiviral Res.* 121, pp. 120–131. DOI: 10.1016/j.antiviral.2015.07.002.

Crossman LC, Gould VC, Dow JM, et al. (2008). "The complete genome, comparative and functional analysis of Stenotrophomonas maltophilia reveals an organism heavily shielded by drug resistance determinants." *Genome Biol.* 9.4, R74. DOI: 10.1186/gb-2008-9-4-r74.

Cusumano-Towner M, Li DY, Tuo S, Krishnan G, and Maslove DM (2013). "A social network of hospital acquired infection built from electronic medical record data." *J. Am. Med. Inform. Assoc.* 20.3, pp. 427–34. DOI: 10.1136/amiajnl-2012-001401.

Dahabreh IJ and Kent DM (2014). "Can the learning health care system be educated with observational data?" *JAMA* 312.2, pp. 129–30. DOI: 10.1001/jama.2014.4364.

Danecek P, Auton A, Abecasis G, et al. (2011). "The variant call format and VCFtools". *Bioinformatics* 27.15, pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330.

Darling ACE, Mau B, Blattner FR, and Perna NT (2004). "Mauve: multiple alignment of conserved genomic sequence with rearrangements." *Genome Res.* 14.7, pp. 1394–403. DOI: 10.1101/gr.2289704.

Darling AE, Mau B, and Perna NT (2010). "Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement". *PLoS One* 5.6. DOI: 10.1371/journal.pone.0011147.

Davies KA, Longshaw CM, Davis GL, et al. (2014). "Underdiagnosis of Clostridium difficile across Europe: The European, multicentre, prospective, biannual, point-prevalence study of Clostridium difficile infection in hospitalised patients with diarrhoea (EUCLID)". *Lancet Infect. Dis.* 14.12, pp. 1208–1219. DOI: 10.1016/S1473-3099(14)70991-0.

Delcher AL, Salzberg SL, and Phillippy AM (2003). "Using MUMmer to identify similar regions in large sequence sets." *Curr. Protoc. Bioinforma.* Chapter 10, Unit 10.3. DOI: 10.1002/0471250953.bi1003s00.

DeLisle S, Kim B, Deepak J, et al. (2013). "Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy." *PLoS One* 8.8, e70944. DOI: 10.1371/journal.pone.0070944.

Deloger M, El Karoui M, and Petit MA (2009). "A genomic distance based on MUM indicates discontinuity between most bacterial species and genera". *J. Bacteriol.* 91.1, pp. 91–99. DOI: 10.1128/JB.01202-08.

Didelot X and Wilson DJ (2015). "ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes". *PLoS Comput. Biol.* 11.2, pp. 1–18. DOI: 10.1371/journal.pcbi.1004041.

Didelot X, Eyre DW, Cule M, et al. (2012a). "Microevolutionary analysis of Clostridium difficile genomes to investigate transmission." *Genome Biol.* 13.12, R118. DOI: 10.1186/gb-2012-13-12-r118.

Didelot X, Bowden R, Wilson DJ, Peto TEA, and Crook DW (2012b). "Transforming clinical microbiology with bacterial genome sequencing". *Nat. Rev. Genet.* 13.9, pp. 601–612. DOI: 10.1038/nrg3226.

Dingle KE, Didelot X, Quan TP, et al. (2017). "Effects of control interventions on Clostridium difficile infection in England: an observational study." *Lancet Infect. Dis.* 17.4, pp. 411–421. DOI: 10.1016/S1473-3099(16)30514-X.

Doern CD (2013). "Integration of technology into clinical practice." *Clin. Lab. Med.* 33.3, pp. 705–29. DOI: 10.1016/j.cll.2013.03.004.

Down TA, Piipari M, and Hubbard TJP (2011). "Dalliance: interactive genome viewing on the web." *Bioinformatics* 27.6, pp. 889–90. DOI: 10.1093/bioinformatics/btr020.

Dreszer TR, Karolchik D, Zweig AS, et al. (2011). "The UCSC Genome Browser database: extensions and updates 2011". *Nucleic Acids Res.* 40.D1, pp. D918–D923. DOI: 10.1093/nar/gkr1055.

Drozd EM, Inocencio TJ, Braithwaite S, et al. (2015). "Mortality, hospital costs, payments, and readmissions associated with Clostridium difficile infection among Medicare beneficiaries". *Infect. Dis. Clin. Pract.* 23.6, pp. 318–323. DOI: 10.1097/IPC.0000000000000299.

Dubberke ER, Reske KA, Olsen MA, McDonald LC, and Fraser VJ (2008). "Short- and Long-Term Attributable Costs of Clostridium difficile-Associated Disease in Nonsurgical Inpatients". *Clin. Infect. Dis.* 46.4, pp. 497–504. DOI: 10.1086/526530.

Dubberke ER, Reske KA, McDonald LC, and Fraser VJ (2006). "ICD-9 codes and surveillance for Clostridium difficile-associated disease". *Emerg. Infect. Dis.* 12.10, pp. 1576–1579. DOI: 10.3201/eid1210.060016.

Dubberke ER, Yan Y, Reske KA, et al. (2011). "Development and validation of a Clostridium difficile infection risk prediction model." *Infect. Control Hosp. Epidemiol.* 32.4, pp. 360–6. DOI: 10.1086/658944.

Dubberke ER, Schaefer E, Reske KA, et al. (2014a). "Attributable inpatient costs of recurrent Clostridium difficile infections". *Infect. Control Hosp. Epidemiol.* 35.11, pp. 1400–1407. DOI: 10.1086/678428.

Dubberke ER, Carling P, Carrico R, et al. (2014b). "Strategies to Prevent Clostridium difficile Infections in Acute Care Hospitals: 2014 Update". *Infect. Control Hosp. Epidemiol.* 35.6, pp. 628–645. DOI: 10.1086/676023.

Eberl M, Roberts GW, Meuter S, et al. (2009). "A rapid crosstalk of human gammadelta T cells and monocytes drives the acute inflammation in bacterial infections". *PLoS Pathog.* 5.2. DOI: 10.1371/journal.ppat.1000308.

Edgell SE and Noon SM (1984). "Effect of violation of normality on the t test of the correlation coefficient." *Psychol. Bull.* 95.3, pp. 576–583. DOI: 10.1037/0033-2909.95.3.576.

Ellery PJ, Tippett E, Chiu YL, et al. (2007). "The CD16+ Monocyte Subset Is More Permissive to Infection and Preferentially Harbors HIV-1 In Vivo". *J. Immunol.* 178.10, pp. 6581–6589. DOI: 10.4049/jimmunol.178.10.6581.

Emilsson V, Thorleifsson G, Zhang B, et al. (2008). "Genetics of gene expression and its effect on disease." *Nature* 452.7186, pp. 423–8. DOI: 10.1038/nature06758.

Etheredge LM (2007). "A rapid-learning health system." *Health Aff. (Millwood).* 26.2, w107–18. DOI: 10.1377/hlthaff.26.2.w107.

Eyre DW, Golubchik T, Gordon NC, et al. (2012). "A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance." *BMJ Open* 2.3. DOI: 10.1136/bmjopen-2012-001124.

Eyre DW, Cule ML, Wilson DJ, et al. (2013). "Diverse sources of C. difficile infection identified on whole-genome sequencing." *N. Engl. J. Med.* 369.13, pp. 1195–1205. DOI: 10.1056/NEJMoa1216064.

Fabregat A, Sidiropoulos K, Garapati P, et al. (2016). "The reactome pathway knowledgebase". *Nucleic Acids Res.* 44.D1, pp. D481–D487. DOI: 10.1093/nar/gkv1351.

Fang G, Munera D, Friedman DI, et al. (2012). "Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing". *Nat. Biotechnol.* 30.12, pp. 1232–1239. DOI: 10.1038/nbt.2432.

Felsen UR, Bellin EY, Cunningham CO, and Zingman BS (2014). "Development of an electronic medical record-based algorithm to identify patients with unknown HIV status." *AIDS Care* 26.10, pp. 1318–25. DOI: 10.1080/09540121.2014.911813.

Fietta AM, Morosini M, Meloni F, Bianco AM, and Pozzi E (2002). "Pharmacological Analysis of Signal Transduction Pathways Required for Mycobacterium Tuberculosis -Induced Il-8 and Mcp-1 Production in Human Peripheral Monocytes". *Cytokine* 19.5, pp. 242–249. DOI: 10.1006/cyto.2002.1968.

Fingerle G, Pforte A, Passlick B, et al. (1993). "The novel subset of CD14+/CD16+ blood monocytes is expanded in sepsis patients." *Blood* 82.10, pp. 3170–6.

Frazee AC, Pertea G, Jaffe AE, et al. (2014). "Flexible analysis of transcriptome assemblies with Ballgown". *bioRxiv*, p. 003665. DOI: 10.1101/003665.

Freese NH, Norris DC, and Loraine AE (2016). "Integrated genome browser: Visual analytics platform for genomics". *Bioinformatics* 32.14, pp. 2089–2095. DOI: 10.1093/bioinformatics/btw069.

Friedman J, Hastie T, and Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *J. Stat. Softw.* 33.1, pp. 1–11. DOI: 10.18637/jss.v033.i01.

Furman D, Jojic V, Kidd B, et al. (2013). "Apoptosis and other immune biomarkers predict influenza vaccine responsiveness." *Mol. Syst. Biol.* 9.659, p. 659. DOI: 10.1038/msb.2013.15.

Gabriel L and Beriot-Mathiot A (2014). "Hospitalization stay and costs attributable to Clostridium difficile infection: A critical review". *J. Hosp. Infect.* 88.1, pp. 12–21. DOI: 10.1016/j.jhin.2014.04.011.

Galo SS, González K, Téllez Y, et al. (2017). "Development of in-house serological methods for diagnosis and surveillance of chikungunya". *Pan Am. J. Public Heal.* 41, e56.

Garrison MW, Anderson DE, Campbell DM, et al. (1996). "Stenotrophomonas maltophilia: Emergence of multidrug-resistant strains during therapy and in an in vitro pharmacodynamic chamber model". *Antimicrob. Agents Chemother.* 40.12, pp. 2859–2864.

Gawande A (2015). "Overkill". *New Yorker* 2015 May 1, pp. 42–53.

Germain RN, Meier-Schellersheim M, Nita-Lazar A, and Fraser IDC (2011). "Systems biology in immunology: a computational modeling perspective." *Annu. Rev. Immunol.* 29, pp. 527–85. DOI: 10.1146/annurev-immunol-030409-101317.

Ghantoji SS, Sail K, Lairson DR, DuPont HL, and Garey KW (2010). "Economic healthcare costs of Clostridium difficile infection: A systematic review". *J. Hosp. Infect.* 74.4, pp. 309–318. DOI: 10.1016/j.jhin.2009.10.016.

Ginsberg J, Mohebbi MH, Patel RS, et al. (2009). "Detecting influenza epidemics using search engine query data." *Nature* 457.7232, pp. 1012–4. DOI: 10.1038/nature07634.

Girdlestone J (1995). "Regulation of HLA Class I Loci by Interferons". *Immunobiology* 193.2–4, pp. 229–237. DOI: 10.1016/S0171-2985(11)80548-6.

Goecks J, Nekrutenko A, and Taylor J (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol.* 11.8, R86. DOI: 10.1186/gb-2010-11-8-r86.

Gordon NC, Price JR, Cole K, et al. (2014). "Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing." *J. Clin. Microbiol.* 52.4, pp. 1182–91. DOI: 10.1128/JCM.03117-13.

Gottesman O, Kuivaniemi H, Tromp G, et al. (2013). "The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future." *Genet. Med.* 15.10, pp. 761–71. DOI: 10.1038/gim.2013.72.

Graves N, Harbarth S, Beyersmann J, et al. (2010). "Estimating the cost of health care-associated infections: mind your p's and q's." *Clin. Infect. Dis.* 50.7, pp. 1017–1021. DOI: 10.1086/651110.

Gray BH, Bowden T, Johansen I, and Koch S (2011). "Electronic health records: an international perspective on "meaningful use"." *Issue Brief (Commonw. Fund)* 28.November, pp. 1–18.

Greco G, Shi W, Michler RE, et al. (2015). "Costs associated with health care-associated infections in cardiac surgery". *J. Am. Coll. Cardiol.* 65.1, pp. 15–23. DOI: 10.1016/j.jacc.2014.09.079.

Halpin T and Morgan T (2010). *Information Modeling and Relational Databases*. 2nd ed. Elsevier Science, p. 976.

Hannenhalli S and Pevzner PA (1999). "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals". *J. ACM* 46.1, pp. 1–27. DOI: 10.1145/300515.300516.

Harris SR, Cartwright EJP, Török ME, et al. (2013). "Whole-genome sequencing for analysis of an outbreak of methicillin-resistant Staphylococcus aureus: a descriptive study." *Lancet Infect. Dis.* 13.2, pp. 130–136. DOI: 10.1016/S1473-3099(12)70268-2.

Haukoos JS and Lewis RJ (2015). "The Propensity Score". *JAMA* 314.15, pp. 1637–8. DOI: 10.1001/jama.2015.13480.

Henry J, Pylypchuk Y, Searcy T, and Patel V (2016). "Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2015". *ONC Data Br.* no.35.

Her Z, Malleret B, Chan M, et al. (2010). "Active infection of human blood monocytes by Chikungunya virus triggers an innate immune response". *J. Immunol.* 184.10, pp. 5903–5913. DOI: 10.4049/jimmunol.0904181.

Hernández A, Maté MJ, Sánchez-Díaz PC, et al. (2009). "Structural and functional analysis of SmeT, the repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF". *J. Biol. Chem.* 284.21, pp. 14428–14438. DOI: 10.1074/jbc.M809221200.

Hilker R, Sickinger C, Pedersen CNS, and Stoye J (2012). "UniMoG-a unifying framework for genomic distance calculation: And sorting based on DCJ". *Bioinformatics* 28.19, pp. 2509–2511. DOI: 10.1093/bioinformatics/bts440.

Hilsenrath P, Eakin C, and Fischer K (2015). "Price-transparency and cost accounting: Challenges for health care organizations in the consumer-driven era". *Inq. (United States)* 52. DOI: 10.1177/0046958015574981.

Hirsch EB and Tam VH (2010). "Detection and treatment options for Klebsiella pneumoniae carbapenemases (KPCs): an emerging cause of multidrug-resistant infection." *J. Antimicrob. Chemother.* 65.6, pp. 1119–25. DOI: 10.1093/jac/dkq108.

Ho DE, Imai K, King G, and Stuart EA (2011). "MatchIt : Nonparametric Preprocessing for Parametric Causal Inference". *J. Stat. Softw.* 42.8, pp. 1–28. DOI: 10.18637/jss.v042.i08.

Hoarau JJ, Jaffar Bandjee MC, Krejbich Trotot P, et al. (2010). "Persistent chronic inflammation and infection by Chikungunya arthritogenic alphavirus in spite of a robust host immune response." *J. Immunol.* 184.10, pp. 5914–27. DOI: 10.4049/jimmunol.0900255.

Huan T, Meng Q, Saleh MA, et al. (2015). "Integrative network analysis reveals molecular mechanisms of blood pressure regulation". *Mol. Syst. Biol.* 11.4, pp. 799–799. DOI: 10.15252/msb.20145399.

Hughes GJ, Nickerson E, Enoch DA, et al. (2013). "Impact of cleaning and other interventions on the reduction of hospital-acquired Clostridium difficile infections in two hospitals in England assessed using a breakpoint model". *J. Hosp. Infect.* 84.3, pp. 227–234. DOI: 10.1016/j.jhin.2012.12.018.

Hunt M, Silva ND, Otto TD, et al. (2015). "Circlator: automated circularization of genome assemblies using long sequencing reads". *Genome Biol.* 16.1, p. 294. DOI: 10.1186/s13059-015-0849-0.

Infectious Diseases Society of America (2010). "The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020." *Clin. Infect. Dis.* 50.8, pp. 1081–3. DOI: 10.1086/652237.

Isniewski MARYFW, Ieszkowski PIK, Agorski BRMZ, et al. (2003). "Development of a Clinical Data Warehouse for Hospital Infection Control". *J. Am. Med. Informatics Assoc.* Pp. 454–463. DOI: 10.1197/jamia.M1299.care.

Jamil HM (2013). "Designing integrated computational biology pipelines visually." *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 10.3, pp. 605–18. DOI: 10.1109/TCBB.2013.69.

Joensen KG, Scheutz F, Lund O, et al. (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of vero-

toxigenic Escherichia coli." *J. Clin. Microbiol.* 52.5, pp. 1501–10. DOI: 10.1128/JCM.03617-13.

Just W (2001). "Computational complexity of multiple sequence alignment with SP-score." *J. Comput. Biol.* 8.6, pp. 615–23. DOI: 10.1089/106652701753307511.

Kam YW, Simarmata D, Chow A, et al. (2012). "Early appearance of neutralizing immunoglobulin G3 antibodies is associated with chikungunya virus clearance and long-term clinical protection". *J. Infect. Dis.* 205.7, pp. 1147–1154. DOI: 10.1093/infdis/jis033.

Karr JR, Sanghvi JC, Macklin DN, et al. (2012). "A whole-cell computational model predicts phenotype from genotype." *Cell* 150.2, pp. 389–401. DOI: 10.1016/j.cell.2012.05.044.

Katz MH (2013). "Pay for preventing (not causing) health care-associated infections." *JAMA Intern. Med.* 173.22, p. 2046. DOI: 10.1001/jamainternmed.2013.9754.

Kelvin AA, Banner D, Silvi G, et al. (2011). "Inflammatory cytokine expression is associated with chikungunya virus resolution and symptom severity." *PLoS Negl. Trop. Dis.* 5.8, e1279. DOI: 10.1371/journal.pntd.0001279.

Kent WJ, Sugnet CW, Furey TS, et al. (2002). "The Human Genome Browser at UCSC". *Genome Res.* 12.6, pp. 996–1006. DOI: 10.1101/gr.229102.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, and Karolchik D (2010). "BigWig and BigBed: enabling browsing of large distributed datasets". *Bioinformatics* 26.17, pp. 2204–2207. DOI: 10.1093/bioinformatics/btq351.

Khanafer N, Voirin N, Barbut F, Kuijper E, and Vanhems P (2015). "Hospital management of Clostridium difficile infection: A review of the literature". *J. Hosp. Infect.* 90.2, pp. 91–101. DOI: 10.1016/j.jhin.2015.02.015.

Kidd BA, Peters LA, Schadt EE, and Dudley JT (2014). "Unifying immunology with informatics and multiscale biology." *Nat. Immunol.* 15.2, pp. 118–27. DOI: 10.1038/ni.2787.

Kim D, Langmead B, and Salzberg SL (2015). "HISAT: a fast spliced aligner with low memory requirements." *Nat. Methods* 12.4, pp. 357–60. DOI: 10.1038/nmeth.3317.

Kleef E van, Green N, Goldenberg SD, et al. (2014). "Excess length of stay and mortality due to Clostridium difficile infection: A multi-state modelling approach". *J. Hosp. Infect.* 88.4, pp. 213–217. DOI: 10.1016/j.jhin.2014.08.008.

Klompas M, Haney G, Church D, et al. (2008). "Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance." *PLoS One* 3.7, e2626. DOI: 10.1371/journal.pone.0002626.

Knight DR, Elliott B, Chang BJ, Perkins TT, and Riley TV (2015). "Diversity and Evolution in the Genome of Clostridium difficile". *Clin. Microbiol. Rev.* 28.3, pp. 721–741. DOI: 10.1128/CMR.00127-14.

Kohane IS, Drazen JM, and Campion EW (2012). "A glimpse of the next 100 years in medicine." *N. Engl. J. Med.* 367.26, pp. 2538–9. DOI: 10.1056/NEJMe1213371.

Köser CU, Ellington MJ, Cartwright EJP, et al. (2012). "Routine use of microbial whole genome sequencing in diagnostic and public health microbiology." *PLoS Pathog.* 8.8, e1002824. DOI: 10.1371/journal.ppat.1002824.

Köster J and Rahmann S (2012). "Snakemake-a scalable bioinformatics workflow engine". *Bioinformatics* 28.19, pp. 2520–2522. DOI: 10.1093/bioinformatics/bts480.

Krumholz HM, Terry SF, and Waldstreicher J (2016). "Data Acquisition, Curation, and Use for a Continuously Learning Health System". *JAMA* 316.16, p. 1669. DOI: 10.1001/jama.2016.12537.

Krzywinski M (2013). "Axes, ticks and grids". *Nat. Publ. Gr.* 10.3, p. 183. DOI: 10.1038/nmeth.2337.

Kullar R, Goff DA, Schulz LT, Fox BC, and Rose WE (2013). "The "epic" challenge of optimizing antimicrobial stewardship: the role of electronic medical records and technology." *Clin. Infect. Dis.* 57.7, pp. 1005–13. DOI: 10.1093/cid/cit318.

Kumar V, Wijmenga C, and Xavier RJ (2014). "Genetics of immune-mediated disorders: From genome-wide association to molecular mechanism". *Curr. Opin. Immunol.* 31, pp. 51–57. DOI: 10.1016/j.coi.2014.09.007.

Kurtz S, Phillippy A, Delcher AL, et al. (2004). "Versatile and open software for comparing large genomes." *Genome Biol.* 5.2, R12. DOI: 10.1186/gb-2004-5-2-r12.

Kutmon M, Riutta A, Nunes N, et al. (2016). "WikiPathways: Capturing the full diversity of pathway knowledge". *Nucleic Acids Res.* 44.D1, pp. D488–D494. DOI: 10.1093/nar/gkv1024.

Kwissa M, Nakaya HI, Onlamoon N, et al. (2014). "Dengue Virus Infection Induces Expansion of a CD14+ CD16+ Monocyte Population that Stimulates Plasmablast Differentiation". *Cell Host Microbe* 16.1, pp. 115–127. DOI: 10.1016/j.chom.2014.06.001.

Labadie K, Larcher T, Joubert C, et al. (2010). "Chikungunya disease in nonhuman primates involves long-term viral persistence in macrophages." *J. Clin. Invest.* 120.3, pp. 894–906. DOI: 10.1172/JCI40104.

Lallemand N (2012). "Health Policy Brief: Reducing Waste in Health Care". *Health Aff.* DOI: 10.1377/hpb2012.23.

Langfelder P and Horvath S (2007). "Eigengene networks for studying the relationships between co-expression modules". *BMC Syst. Biol.* 1.1, p. 54. DOI: 10.1186/1752-0509-1-54.

Langfelder P, Zhang B, and Horvath S (2008). "Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R". *Bioinformatics* 24.5, pp. 719–720. DOI: 10.1093/bioinformatics/btm563.

Law CW, Chen Y, Shi W, and Smyth GK (2014). "voom: precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biol.* 15.2, R29. DOI: 10.1186/gb-2014-15-2-r29.

Lazer D, Kennedy R, King G, and Vespignani A (2014). "Big data. The parable of Google Flu: traps in big data analysis." *Science* 343.6176, pp. 1203–5. DOI: 10.1126/science.1248506.

Lee HC, Kosoy R, Becker CE, Dudley JT, and Kidd BA (2017). "Automated cell type discovery and classification through knowledge transfer". *Bioinformatics* 45.2, pp. 846–860. DOI: 10.1093/bioinformatics/btx054.

Leffler DA and Lamont JT (2015). "Clostridium difficile." *N. Engl. J. Med.* 372, pp. 1539–48. DOI: 10.1056/NEJMra1403772.

Leibovici L, Gitelman V, Yehezkelli Y, et al. (1997). "Improving empirical antibiotic treatment: prospective, nonintervention testing of a decision support system." *J. Intern. Med.* 242.5, pp. 395–400.

Lemey P, Rambaut A, Bedford T, et al. (2014). "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2." *PLoS Pathog.* 10.2, e1003932. DOI: 10.1371/journal.ppat.1003932.

Lengauer T and Sing T (2006). "Bioinformatics-assisted anti-HIV therapy." *Nat. Rev. Microbiol.* 4.10, pp. 790–797. DOI: 10.1038/nrmicro1477.

Levine JH, Simonds EF, Bendall SC, et al. (2015). "Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis". *Cell* 162.1, pp. 184–197. DOI: 10.1016/j.cell.2015.05.047.

Li H (2011). "Tabix: fast retrieval of sequence features from generic TAB-delimited files". *Bioinformatics* 27.5, pp. 718–719. DOI: 10.1093/bioinformatics/btq671.

Li H and Durbin R (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform". *Bioinformatics* 26.5, pp. 589–595. DOI: 10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, et al. (2009). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li YP, Propert KJ, and Rosenbaum PR (2001). "Balanced Risk Set Matching". *J. Am. Stat. Assoc.* 96.October 2014, pp. 870–882. DOI: 10.1198/016214501753208573.

Liao KP, Cai T, Savova GK, et al. (2015). "Development of phenotype algorithms using electronic medical records and incorporating natural language processing". *BMJ* 350.apr24 11, h1885–h1885. DOI: 10.1136/bmj.h1885.

Lin Y and Moret BME (2008). "Estimating true evolutionary distances under the DCJ model". *Bioinformatics* 24.13, pp. 114–122. DOI: 10.1093/bioinformatics/btn148.

Linderman MD, Brandt T, Edelmann L, et al. (2014). "Analytical validation of whole exome and whole genome sequencing for clinical applications". *BMC Med. Genomics* 7.1, p. 20. DOI: 10.1186/1755-8794-7-20.

Linderman MD, Bashir A, Diaz GA, et al. (2015). "Preparing the next generation of genomicists: a laboratory-style course in medical genomics". *BMC Med. Genomics* 8.1, pp. 1–3. DOI: 10.1186/s12920-015-0124-y.

Liu M, Deora R, Simons RW, et al. (2004). "Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements". *Nature* 431.7007, pp. 476–481. DOI: 10.1038/nature02844.1..

Liu Q, Qiao C, Marjuki H, et al. (2011). "Combination of PB2 271A and SR Polymorphism at Positions 590/591 Is Critical for Viral Replication and Virulence of Swine Influenza Virus in Cultured Cells and In Vivo". *J. Virol.* 86.2, pp. 1233–1237. DOI: 10.1128/JVI.05699-11.

Lofgren ET, Cole SR, Weber DJ, Anderson DJ, and Moehring RW (2014). "Hospital-Acquired Clostridium difficile Infections". *Epidemiology* 25.4, pp. 570–575. DOI: 10.1097/EDE.0000000000000119.

Longtin Y, Paquet-Bolduc B, Gilca R, et al. (2016). "Effect of Detecting and Isolating Clostridium difficile Carriers at Hospital Admission on the Incidence of C difficile Infections: A Quasi-Experimental Controlled Study." *JAMA Intern. Med.* 176.6, pp. 796–804. DOI: 10.1001/jamainternmed.2016.0177.

Luksza M and Lässig M (2014). "A predictive fitness model for influenza." *Nature* 507.7490, pp. 57–61. DOI: doi:10.1038/nature13087.

Lum FM and Ng LF (2015). "Cellular and molecular mechanisms of chikungunya pathogenesis". *Antiviral Res.* 120, pp. 165–174. DOI: 10.1016/j.antiviral.2015.06.009.

Luo W and Brouwer C (2013). "Pathview: An R/Bioconductor package for pathway-based data integration and visualization". *Bioinformatics* 29.14, pp. 1830–1831. DOI: 10.1093/bioinformatics/btt285.

Luster AD and Ravetch JV (1987). "Biochemical characterization of a gamma interferon-inducible cytokine (IP-10)." *J Exp Med.* 166.4, pp. 1084–97. DOI: 10.1084/jem.166.4.1084.

Mahalingam S, Herring BL, and Halstead SB (2013). "Call to action for dengue vaccine failure." *Emerg. Infect. Dis.* 19.8, pp. 1335–7. DOI: 10.3201/eid1908.121864.

Mandl KD, Kohane IS, McFadden D, et al. (2014). "Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture." *J. Am. Med. Inform. Assoc.* 21.4, pp. 615–20. DOI: 10.1136/amiajnl-2014-002727.

Matsuoka H, Hirooka K, and Fujita Y (2007). "Organization and function of the YsiA regulon of Bacillus subtilis involved in fatty acid degradation". *J. Biol. Chem.* 282.8, pp. 5180–5194. DOI: 10.1074/jbc.M606831200.

McAdam PR, Templeton KE, Edwards GF, et al. (2012). "Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus." *Proc. Natl. Acad. Sci. U. S. A.* 109.23, pp. 9107–12. DOI: 10.1073/pnas.1202869109.

McGlone SM, Bailey RR, Zimmer SM, et al. (2012). "The economic burden of Clostridium difficile". *Clin. Microbiol. Infect.* 18.3, pp. 282–289. DOI: 10.1111/j.1469-0691.2011.03571.x.

McKellar MR and Fendrick AM (2014). "Innovation of novel antibiotics: an economic perspective." *Clin. Infect. Dis.* 59.Suppl 3, S104–7. DOI: 10.1093/cid/ciu530.

McKenna A, Hanna M, Banks E, et al. (2010). "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data". *Genome Res.* 20.9, pp. 1297–1303. DOI: 10.1101/gr.107524.110.

Mei HE, Leipold MD, and Maecker HT (2016). "Platinum-conjugated antibodies for application in mass cytometry". *Cytom. Part A* 89.3, pp. 292–300. DOI: 10.1002/cyto.a.22778.

Mejias A and Ramilo O (2014). "Transcriptional profiling in infectious diseases: ready for prime time?" *J. Infect.* 68 Suppl 1, S94–9. DOI: 10.1016/j.jinf.2013.09.018.

Mi H, Muruganujan A, and Thomas PD (2013). "PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees". *Nucleic Acids Res.* 41.D1, pp. 377–386. DOI: 10.1093/nar/gks1118.

Miner JJ, Yeang HXA, Fox JM, et al. (2015). "Brief report: Chikungunya viral arthritis in the United States: A mimic of seronegative rheumatoid arthritis". *Arthritis Rheumatol.* 67.5, pp. 1214–1220. DOI: 10.1002/art.39027.

Minot S, Grunberg S, Wu GD, Lewis JD, and Bushman FD (2012). "Hypervariable loci in the human gut virome". *Proc. Natl. Acad. Sci.* 109.10, pp. 3962–3966. DOI: 10.1073/pnas.1119061109.

Mitchell BG, Gardner A, Barnett AG, Hiller JE, and Graves N (2014). "The prolongation of length of stay because of Clostridium difficile infection". *Am. J. Infect. Control* 42.2, pp. 164–167. DOI: 10.1016/j.ajic.2013.07.006.

Moehring RW, Lofgren ET, and Anderson DJ (2013). "Impact of Change to Molecular Testing for Clostridium difficile Infection on Healthcare Facility-Associated Incidence Rates." *Infect. Control Hosp. Epidemiol.* 34.10, pp. 1055–61. DOI: 10.1086/673144.

Mohammed EA, Far BH, and Naugler C (2014). "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends." *BioData Min.* 7.1, p. 22. DOI: 10.1186/1756-0381-7-22.

Mwangi MM, Wu SW, Zhou Y, et al. (2007). "Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing." *Proc. Natl. Acad. Sci. U. S. A.* 104.17, pp. 9451–9456. DOI: 10.1073/pnas.0609839104.

Naccache SN, Peggs KS, Mattes FM, et al. (2015). "Diagnosis of Neuroinvasive Astrovirus Infection in an Immunocompromised Adult With Encephalitis by Unbiased Next-Generation Sequencing". *Clin. Infect. Dis.* 60, pp. 919–923. DOI: 10.1093/cid/ciu912.

Naccache SN, Federman S, Veeraraghavan N, et al. (2014). "A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples". *Genome Res.* 24.7, pp. 1180–1192. DOI: 10.1101/gr.171934.113.

Nagar R, Yuan Q, Freifeld CC, et al. (2014). "A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives." en. *J. Med. Internet Res.* 16.10, e236. DOI: 10.2196/jmir.3416.

Nagel AC, Tsou MH, Spitzberg BH, et al. (2013). "The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets." en. *J. Med. Internet Res.* 15.10, e237. DOI: 10.2196/jmir.2705.

Nakaya HI, Gardner J, Poo YS, et al. (2012). "Gene profiling of chikungunya virus arthritis in a mouse model reveals significant overlap with rheumatoid arthritis". *Arthritis Rheum.* 64.11, pp. 3553–3563. DOI: 10.1002/art.34631.

Nasci RS (2014). "Movement of chikungunya virus into the Western hemisphere." *Emerg. Infect. Dis.* 20.8, pp. 1394–5. DOI: 10.3201/eid2008.140333.

Nefzger MD and Drasgow J (1957). "The needless assumption of normality in Pearson's r." *Am. Psychol.* 12.10, pp. 623–625. DOI: 10.1037/h0048216.

Ng LFP, Chow A, Sun YJ, et al. (2009). "IL-1beta, IL-6, and RANTES as biomarkers of Chikungunya severity." *PLoS One* 4.1, e4261. DOI: 10.1371/journal.pone.0004261.

Norton B and Naggie S (2014). "The clinical management of HCV in the HIV-infected patient." *Antivir. Ther.* DOI: 10.3851/IMP2910.

Ogata H, Goto S, Sato K, et al. (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes". *Nucleic Acids Res.* 27.1, pp. 29–34. DOI: 10.1093/nar/27.1.29.

OhAinle M, Balmaseda A, Macalalad AR, et al. (2011). "Dynamics of dengue disease severity determined by the interplay between viral genetics and serotype-specific immunity." *Sci. Transl. Med.* 3.114, 114ra128. DOI: 10.1126/scitranslmed.3003084.

Overbeek R, Olson R, Pusch GD, et al. (2014). "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)." *Nucleic Acids Res.* 42.Database issue, pp. D206–14. DOI: 10.1093/nar/gkt1226.

Pak TR and Kasarskis A (2015). "How Next-Generation Sequencing and Multiscale Data Analysis Will Transform Infectious Disease Management". *Clin. Infect. Dis.* 61.11, pp. 1695–1702. DOI: 10.1093/cid/civ670.

Pak TR and Roth FP (2013). "ChromoZoom: a flexible, fluid, web-based genome browser." *Bioinformatics* 29.3, pp. 384–6. DOI: 10.1093/bioinformatics/bts695.

Pak TR, Altman DR, Attie O, et al. (2015). "Whole-Genome Sequencing Identifies Emergence of a Quinolone Resistance Mutation in a Case of Stenotrophomonas maltophilia Bacteremia." *Antimicrob. Agents Chemother.* 59.11, pp. 7117–20. DOI: 10.1128/AAC.01723-15.

Pathak J, Kho AN, and Denny JC (2013). "Electronic health records-driven phenotyping: challenges, recent advances, and perspectives." *J. Am. Med. Inform. Assoc.* 20.e2, e206–11. DOI: 10.1136/amiajnl-2013-002428.

Paulson LD (2005). "Web Applications with Ajax". *IEEE Comput.* 38.10, pp. 14–17. DOI: 10.1109/MC.2005.330.

Peaper DR, Havill NL, Aniskiewicz M, et al. (2015). "Pseudo-outbreak of Actinomyces graevenitzii associated with bronchoscopy". *J. Clin. Microbiol.* 53.1, pp. 113–117. DOI: 10.1128/JCM.02302-14.

Pertea M, Pertea GM, Antonescu CM, et al. (2015). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." *Nat. Biotechnol.* 33.3, pp. 290–5. DOI: 10.1038/nbt.3122.

Pertea M, Kim D, Pertea GM, Leek JT, and Salzberg SL (2016). "Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown". *Nat Protoc.* 11.9, pp. 1650–1667. DOI: 10.1038/nprot.2016.095.

Pimentel HJ, Bray N, Puente S, Melsted P, and Pachter L (2016). "Differential analysis of RNA-Seq incorporating quantification uncertainty". *bioRxiv*, p. 058164. DOI: 10.1101/058164.

Plante KS, Rossi SL, Bergren NA, Seymour RL, and Weaver SC (2015). "Extended Preclinical Safety, Efficacy and Stability Testing of a Live-attenuated Chikungunya Vaccine Candidate". *PLoS Negl. Trop. Dis.* 9.9, e0004007. DOI: 10.1371/journal.pntd.0004007.

Polage CR, Gyorke CE, Kennedy MA, et al. (2015). "Overdiagnosis of Clostridium difficile Infection in the Molecular Test Era." *JAMA Intern. Med.* 175.11, pp. 1–10. DOI: 10.1001/jamainternmed.2015.4114.

Poland GA, Ovsyannikova IG, Kennedy RB, Lambert ND, and Kirkland JL (2014). "A systems biology approach to the effect of aging, immunosenescence and vaccine response". *Curr. Opin. Immunol.* 29, pp. 62–68. DOI: 10.1016/j.coi.2014.04.005.

Poo YS, Rudd PA, Gardner J, et al. (2014). "Multiple Immune Factors Are Involved in Controlling Acute and Chronic Chikungunya Virus Infection". *PLoS Negl. Trop. Dis.* 8.12, e3354. DOI: 10.1371/journal.pntd.0003354.

Price JR, Golubchik T, Cole K, et al. (2014). "Whole-genome sequencing shows that patient-to-patient transmission rarely accounts for acquisition of Staphylococcus aureus in an intensive care unit." *Clin. Infect. Dis.* 58.5, pp. 609–18. DOI: 10.1093/cid/cit807.

Querec TD, Akondy RS, Lee EK, et al. (2009). "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans." *Nat. Immunol.* 10.1, pp. 116–25. DOI: 10.1038/ni.1688.

Raj A, Dewar M, Palacios G, Rabadan R, and Wiggins CH (2011). "Identifying hosts of families of viruses: a machine learning approach." *PLoS One* 6.12, e27631. DOI: 10.1371/journal.pone.0027631.

Ram D, Leshkowitz D, Gonzalez D, et al. (2015). "Evaluation of GS Junior and MiSeq next-generation sequencing technologies as an alternative to Trugene

population sequencing in the clinical HIV laboratory". *J. Virol. Methods* 212, pp. 12–16. DOI: 10.1016/j.jviromet.2014.11.003.

Rasko DA, Webster DR, Sahl JW, et al. (2011). "Origins of the E. Coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany". *N. Engl. J. Med.* 365.8, pp. 709–717. DOI: 10.1056/NEJMoa1106920.

Rea S, Pathak J, Savova G, et al. (2012). "Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project". *J. Biomed. Inform.* 45.4, pp. 763–771. DOI: 10.1016/j.jbi.2012.01.009.

Read TD and Massey RC (2014). "Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology". *Genome Med.* 6.11, p. 109. DOI: 10.1186/s13073-014-0109-z.

Reis BY and Mandl KD (2003). "Integrating syndromic surveillance data across multiple locations: effects on outbreak detection performance." *AMIA Annu. Symp. Proc.* Pp. 549–53.

Reuter S, Hunt M, Peacock SJ, et al. (2016). "Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology". *Microb. Genomics* 2.9, pp. 1689–1699. DOI: 10.1099/mgen.0.000085.

Rhee C, Murphy MV, Li L, Platt R, and Klompas M (2015). "Improving documentation and coding for acute organ dysfunction biases estimates of changing sepsis severity and burden: a retrospective study". *Crit. Care* 19.1, pp. 1–11. DOI: 10.1186/s13054-015-1048-9.

Ritchie ME, Phipson B, Wu D, et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Res.* 43.7, e47–e47. DOI: 10.1093/nar/gkv007.

Roach DJ, Burton JN, Lee C, et al. (2015). "A Year of Infection in the Intensive Care Unit: Prospective Whole Genome Sequencing of Bacterial Clinical Isolates Reveals Cryptic Transmissions and Novel Microbiota". *PLOS Genet.* 11.7, e1005413. DOI: 10.1371/journal.pgen.1005413.

Robert X and Gouet P (2014). "Deciphering key features in protein structures with the new ENDscript server". *Nucleic Acids Res.* 42.April, pp. 320–324. DOI: 10.1093/nar/gku316.

Rodrigues R, Barber GE, and Ananthakrishnan AN (2016). "A Comprehensive Study of Costs Associated With Recurrent Clostridium difficile Infection". *Infect. Control Hosp. Epidemiol.* Pp. 1–7. DOI: 10.1017/ice.2016.246.

Rogacev KS, Cremers B, Zawada AM, et al. (2012). "CD14++CD16+ monocytes independently predict cardiovascular events: A cohort study of 951 patients referred for elective coronary angiography". *J. Am. Coll. Cardiol.* 60.16, pp. 1512–1520. DOI: 10.1016/j.jacc.2012.07.019.

Rolph MS, Foo SS, and Mahalingam S (2015). "Emergent chikungunya virus and arthritis in the Americas". *Lancet Infect. Dis.* 15.9, pp. 1007–1008. DOI: 10.1016/S1473-3099(15)00231-5.

Romano PS and Mark DH (1994). "Bias in the coding of hospital discharge data and its implications for quality assessment." *Med. Care* 32.1, pp. 81–90.

Rulli NE, Rolph MS, Srikiatkhachorn A, et al. (2011). "Protection from arthritis and myositis in a mouse model of acute chikungunya virus disease by bindarit, an inhibitor of monocyte chemotactic protein-1 synthesis". *J. Infect. Dis.* 204.7, pp. 1026–1030. DOI: 10.1093/infdis/jir470.

Samusik N, Good Z, Spitzer MH, Davis KL, and Nolan GP (2016). "Automated mapping of phenotype space with single-cell data." *Nat. Methods* 13.6, pp. 493–496. DOI: 10.1038/nmeth.3863.

Sánchez P, Alonso A, and Martinez JL (2002). "Cloning and characterization of SmeT, a repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF". *Antimicrob. Agents Chemother.* 46.11, pp. 3386–3393. DOI: 10.1128/AAC.46.11.3386-3393.2002.

Sanchez P, Alonso A, and Martinez JL (2004). "Regulatory regions of smeDEF in Stenotrophomonas maltophilia strains expressing different amounts of the multidrug efflux pump SmeDEF". *Antimicrob. Agents Chemother.* 48.6, pp. 2274–2276. DOI: 10.1128/AAC.48.6.2274-2276.2004.

Saraiva M and O'Garra A (2010). "The regulation of IL-10 production by immune cells". *Nat. Rev. Immunol.* 10.3, pp. 170–181. DOI: 10.1038/nri2711.

Scheurer DB, Hicks LS, Cook EF, and Schnipper JL (2007). "Accuracy of ICD-9 coding for Clostridium difficile infections: a retrospective cohort." *Epidemiol. Infect.* 135.6, pp. 1010–3. DOI: 10.1017/S0950268806007655.

Schilte C, Couderc T, Chretien F, et al. (2010). "Type I IFN controls chikungunya virus via its action on nonhematopoietic cells." *J. Exp. Med.* 207.2, pp. 429–442. DOI: 10.1084/jem.20090851.

Schilte C, Buckwalter MR, Laird ME, et al. (2012). "Cutting edge: independent roles for IRF-3 and IRF-7 in hematopoietic and nonhematopoietic cells during host response to Chikungunya infection." *J. Immunol.* 188.7, pp. 2967–71. DOI: 10.4049/jimmunol.1103185.

Schilte C, Staikovsky F, Couderc T, et al. (2013). "Chikungunya Virus-associated Long-term Arthralgia: A 36-month Prospective Longitudinal Study". *PLoS Negl. Trop. Dis.* 7.3. DOI: 10.1371/journal.pntd.0002137.

Scott RD (2009). *The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention*.

Seemann T (2014). "Prokka: Rapid prokaryotic genome annotation". *Bioinformatics* 30.14, pp. 2068–2069. DOI: 10.1093/bioinformatics/btu153.

Sen N, Mukherjee G, and Arvin AM (2015). "Single cell mass cytometry reveals remodeling of human T cell phenotypes by varicella zoster virus". *Methods* 90, pp. 85–94. DOI: 10.1016/j.ymeth.2015.07.008.

Serbina NV, Jia T, Hohl TM, and Pamer EG (2008). "Monocyte-mediated defense against microbial pathogens". *Annu. Rev. Immunol.* 26, pp. 421–52. DOI: 10.1146/annurev.immunol.26.021607.090326.

Shortell SM, Colla CH, Lewis VA, et al. (2015). "Accountable Care Organizations: The National Landscape." *J. Health Polit. Policy Law* 40.4, pp. 647–668. DOI: 10.1215/03616878-3149976.

Sievers F, Wilm A, Dineen D, et al. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". *Mol. Syst. Biol.* 7.539. DOI: 10.1038/msb.2011.75.

Silva JC, Shah SC, Rumoro DP, et al. (2013). "Comparing the accuracy of syndrome surveillance systems in detecting influenza-like illness: GUARDIAN vs. RODS vs. electronic medical record reports". *Artif. Intell. Med.* 59.3, pp. 169–174. DOI: 10.1016/j.artmed.2013.09.001.

Sim X, Jensen RA, Ikram MK, et al. (2013). "Genetic Loci for Retinal Arteriolar Microcirculation". *PLoS One* 8.6, pp. 1–12. DOI: 10.1371/journal.pone.0065804.

Sing T, Sander O, Beerenwinkel N, and Lengauer T (2005). "ROCR: Visualizing classifier performance in R". *Bioinformatics* 21.20, pp. 3940–3941. DOI: 10.1093/bioinformatics/bti623.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, and Holmes IH (2009). "JBrowse: A next-generation genome browser". *Genome Res.* 19.9, pp. 1630–1638. DOI: 10.1101/gr.094607.109.

Skrzeczyńska-Moncznik J, Bzowska M, Loseke S, et al. (2008). "Peripheral blood CD14high CD16+ monocytes are main producers of IL-10". *Scand. J. Immunol.* 67.2, pp. 152–159. DOI: 10.1111/j.1365-3083.2007.02051.x.

Smoot ME, Ono K, Ruscheinski J, Wang PL, and Ideker T (2011). "Cytoscape 2.8: new features for data integration and network visualization." *Bioinformatics* 27.3, pp. 431–2. DOI: 10.1093/bioinformatics/btq675.

Snow J (1855). *On the Mode of Communication of Cholera*. 2nd ed. London, England: John Churchill.

Sohn JI and Nam JW (2016). "The present and future of de novo whole-genome assembly". *Brief. Bioinform.* January 2015, bbw096. DOI: 10.1093/bib/bbw096.

Sourisseau M, Schilte C, Casartelli N, et al. (2007). "Characterization of reemerging chikungunya virus." *PLoS Pathog.* 3.6, e89. DOI: 10.1371/journal.ppat.0030089.

Srikhanta YN, Fox KL, and Jennings MP (2010). "The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes." *Nat. Rev. Microbiol.* 8.3, pp. 196–206. DOI: 10.1038/nrmicro2283.

Stalker J, Gibbins B, Meidl P, et al. (2004). "The Ensembl Web site: mechanics of a genome browser." *Genome Res.* 14.5, pp. 951–5. DOI: 10.1101/gr.1863004.

Stamatakis A, Ludwig T, and Meier H (2005). "RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees." *Bioinformatics* 21.4, pp. 456–63. DOI: 10.1093/bioinformatics/bti191.

Stamatakis A (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30.9, pp. 1312–3. DOI: 10.1093/bioinformatics/btu033.

Stansfield BK and Ingram DA (2015). "Clinical significance of monocyte heterogeneity." *Clin. Transl. Med.* 4, p. 5. DOI: 10.1186/s40169-014-0040-3.

Stein LD (2002). "The Generic Genome Browser: A Building Block for a Model Organism System Database". *Genome Res.* 12.10, pp. 1599–1610. DOI: 10.1101/gr.403602.

Stein RA (2011). "Super-spreaders in infectious diseases". *Int. J. Infect. Dis.* 15.8, e510–e513. DOI: 10.1016/j.ijid.2010.06.020.

Stevens VW, Khader K, Nelson RE, et al. (2015). "Excess Length of Stay Attributable to Clostridium difficile Infection (CDI) in the Acute Care Setting: A Multistate Model." *Infect. Control Hosp. Epidemiol.* 36.Cdi, pp. 1–7. DOI: 10.1017/ice.2015.132.

Suhrbier A, Jaffar-Bandjee MC, and Gasque P (2012). "Arthritogenic alphaviruses: an overview". *Nat. Rev. Rheumatol.* 8.7, pp. 420–429. DOI: 10.1038/nrrheum.2012.64.

Tanner J, Khan D, Anthony D, and Paton J (2009). "Waterlow score to predict patients at risk of developing Clostridium difficile-associated disease". *J. Hosp. Infect.* 71.3, pp. 239–244. DOI: 10.1016/j.jhin.2008.11.017.

Teng TS, Kam YW, Lee B, et al. (2015). "A Systematic Meta-analysis of Immune Signatures in Patients With Acute Chikungunya Virus Infection". *J. Infect. Dis.* 211, pp. 1925–1935. DOI: 10.1093/infdis/jiv049.

Teng TSS, Foo SSS, Simamarta D, et al. (2012). "Viperin restricts chikungunya virus replication and pathology". *J. Clin. Invest.* 122.12, pp. 4447–4460. DOI: 10.1172/JCI63120.

Tesler G (2002). "GRIMM: genome rearrangements web server." *Bioinformatics* 18.3, pp. 492–493. DOI: 10.1093/bioinformatics/18.3.492.

The 1000 Genomes Project Consortium (2010). "A map of human genome variation from population-scale sequencing." *Nature* 467.7319, pp. 1061–73. DOI: 10.1038/nature09534.

The Gene Ontology Consortium (2015). "Gene ontology consortium: Going forward". *Nucleic Acids Res.* 43.D1, pp. D1049–D1056. DOI: 10.1093/nar/gku1179.

Thorvaldsdóttir H, Robinson JT, and Mesirov JP (2013). "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration". *Brief. Bioinform.* 14.2, pp. 178–192. DOI: 10.1093/bib/bbs017.

Tong SY, Holden MT, Nickerson EK, et al. (2015). "Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting". *Genome Res.* 25.1, pp. 111–118. DOI: 10.1101/gr.174730.114.

Trapnell C, Hendrickson DG, Sauvageau M, et al. (2013). "Differential analysis of gene regulation at transcript resolution with RNA-seq." *Nat. Biotechnol.* 31.1, pp. 46–53. DOI: 10.1038/nbt.2450.

Treangen TJ, Ondov BD, Koren S, and Phillippy AM (2014). "The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes." *Genome Biol.* 15.11, p. 524. DOI: 10.1186/PREACCEPT-2573980311437212.

Tsukamoto M, Seta N, Yoshimoto K, et al. (2017). "CD14(bright) CD16+ intermediate monocytes are induced by interleukin-10 and positively correlate with disease activity in rheumatoid arthritis." *Arthritis Res. Ther.* 19.1, p. 28. DOI: 10.1186/s13075-016-1216-6.

Tufte ER (2001). *The Visual Display of Quantitative Information.* 2nd editio. Cheshire, CT: Graphics Press.

Turon A (2013). "Understanding and Expressing Scalable Concurrency". PhD thesis. Northeastern University.

Valdezate S, Vindel A, Echeita A, Baquero F, and Cantón R (2002). "Topoisomerase II and IV quinolone resistance determining regions in Stenotrophomonas maltophilia clinical isolates with different levels of quinolone susceptibility". *Antimicrob. Agents Chemother.* 46.3, pp. 665–671. DOI: 10.1128/AAC.46.3.665-671.2002.

Valdezate S, Vindel A, Saéz-Nieto JA, Baquero F, and Cantón R (2005). "Preservation of topoisomerase genetic sequences during in vivo and in vitro development of high-level resistance to ciprofloxacin in isogenic Stenotrophomonas maltophilia strains". *J. Antimicrob. Chemother.* 56.1, pp. 220–223. DOI: 10.1093/jac/dki182.

Van Damme J, Proost P, Put W, et al. (1994). "Induction of monocyte chemotactic proteins MCP-1 and MCP-2 in human fibroblasts and leukocytes by cytokines and cytokine inducers. Chemical synthesis of MCP-2 and development of a specific RIA." *J. Immunol.* 152.11, pp. 5495–502. DOI: 10.4049/jimmunol.1103138.

Van der Auwera GA, Carneiro MO, Hartl C, et al. (2013). "From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline". *Curr. Protoc. Bioinforma.* SUPL.43, pp. 1–33. DOI: 10.1002/0471250953.bi1110s43.

Vanderkam D, Aksoy BA, Hodes I, Perrone J, and Hammerbacher J (2016). "pileup.js: a JavaScript library for interactive and in-browser visualization of genomic data". *Bioinformatics* 32.15, pp. 2378–2379. DOI: 10.1093/bioinformatics/btw167.

Veiga-Castelli LC, Rosa e Silva JC, Meola J, et al. (2010). "Genomic alterations detected by comparative genomic hybridization in ovarian endometriomas". *Brazilian J. Med. Biol. Res.* 43.8, pp. 799–805. DOI: 10.1590/S0100-879X2010007500072.

Vincent L (2007). "Taking online maps down to street level". *Computer (Long. Beach. Calif).* 40.12, pp. 118–120. DOI: 10.1109/MC.2007.442.

Volk SM, Chen R, Tsetsarkin KA, et al. (2010). "Genome-Scale Phylogenetic Analyses of Chikungunya Virus Reveal Independent Emergences of Recent

Epidemics and Various Evolutionary Rates". *J. Virol.* 84.13, pp. 6497–6504. DOI: 10.1128/JVI.01603-09.

Waggoner JJ, Gresh L, Vargas MJ, et al. (2016). "Viremia and Clinical Presentation in Nicaraguan Patients Infected with Zika Virus, Chikungunya Virus, and Dengue Virus." *Clin. Infect. Dis.* 63, pp. 1–7. DOI: 10.1093/cid/ciw589.

Wagner B, Filice GA, Drekonja D, et al. (2014). "Antimicrobial stewardship programs in inpatient hospital settings: a systematic review." *Infect. Control Hosp. Epidemiol.* 35.10, pp. 1209–28. DOI: 10.1086/678057.

Wang YL, Scipione MR, Dubrovskaya Y, and Papadopoulos J (2014). "Monotherapy with fluoroquinolone or trimethoprim-sulfamethoxazole for treatment of Stenotrophomonas maltophilia infections". *Antimicrob. Agents Chemother.* 58.1, pp. 176–182. DOI: 10.1128/AAC.01751-13.

Wasmuth EV and Lima CD (2016). "UniProt: the universal protein knowledgebase". *Nucleic Acids Res.* 45.November 2016, pp. 1–12. DOI: 10.1093/nar/gkw1152.

Wauquier N, Becquart P, Nkoghe D, et al. (2011). "The acute phase of Chikungunya virus infection in humans is associated with strong innate immunity and T CD8 cell activation". *J. Infect. Dis.* 204.1, pp. 115–123. DOI: 10.1093/infdis/jiq006.

Weaver SC and Lecuit M (2015). "Chikungunya virus and the global spread of a mosquito-borne disease." *N. Engl. J. Med.* 372.13, pp. 1231–9. DOI: 10.1056/NEJMra1406035.

Weger-Lucarelli J, Chu H, Aliota MT, Partidos CD, and Osorio JE (2014). "A Novel MVA Vectored Chikungunya Virus Vaccine Elicits Protective Immunity in Mice." *PLoS Negl. Trop. Dis.* 8.7, e2970. DOI: 10.1371/journal.pntd.0002970.

Wheeler DL, Church DM, Federhen S, et al. (2003). "Database resources of the national center for biotechnology". *Nucleic Acids Res.* 31.1, pp. 28–33. DOI: 10.1093/nar/gkg033.

Wiens J, Guttag J, and Horvitz E (2014). "A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions." *J. Am. Med. Inform. Assoc.* 21.4, pp. 699–706. DOI: 10.1136/amiajnl-2013-002162.

Wilcox MH, Cunniffe JG, Trundle C, and Redpath C (1996). "Financial burden of hospital-acquired Clostridium difficile infection." *J. Hosp. Infect.* 34.1, pp. 23–30. DOI: 10.1016/S0195-6701(96)90122-X.

Wilson JAC, Prow NA, Schroder WA, et al. (2017). "RNA-Seq analysis of chikungunya virus infection and identification of granzyme A as a major promoter of arthritic inflammation." *PLoS Pathog.* 13.2, e1006155. DOI: 10.1371/journal.ppat.1006155.

Wilson MR, Naccache SN, Samayoa E, et al. (2014). "Actionable diagnosis of neuroleptospirosis by next-generation sequencing." *N. Engl. J. Med.* 370.25, pp. 2408–17. DOI: 10.1056/NEJMoa1401268.

Wong DDY, Choy KT, Chan RWY, et al. (2012). "Comparable Fitness and Transmissibility between Oseltamivir-Resistant Pandemic 2009 and Seasonal H1N1 Influenza Viruses with the H275Y Neuraminidase Mutation". *J. Virol.* 86.19, pp. 10558–10570. DOI: 10.1128/JVI.00985-12.

Wong KL, Tai JJY, Wong WC, et al. (2011). "Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets." *Blood* 118.5, e16–31. DOI: 10.1182/blood-2010-12-326355.

Zawada AM, Rogacev KS, Rotter B, et al. (2011). "SuperSAGE evidence for CD14++CD16+ monocytes as a third monocyte subset." *Blood* 118.12, e50–61. DOI: 10.1182/blood-2011-01-326827.

Zazzi M, Incardona F, Rosen-Zvi M, et al. (2012). "Predicting Response to Antiretroviral Treatment by Machine Learning: The EuResist Project". *Intervirology* 55.2, pp. 123–127. DOI: 10.1159/000332008.

Zhang B and Horvath S (2005). "A general framework for weighted gene co-expression network analysis." *Stat. Appl. Genet. Mol. Biol.* 4.1, Article17. DOI: 10.2202/1544-6115.1128.

Zhang S, Palazuelos-Munoz S, Balsells EM, et al. (2016). "Cost of hospital management of Clostridium difficile infection in United States—a meta-analysis and modelling study". *BMC Infect. Dis.* 16.1, p. 447. DOI: 10.1186/s12879-016-1786-6.

Zhang Y, Zhu J, Li Y, et al. (2013). "Glycosylation on Hemagglutinin Affects the Virulence and Pathogenicity of Pandemic H1N1/2009 Influenza A Virus in Mice". *PLoS One* 8.4, e61397.

Zhao Y, Niu W, Sun Y, et al. (2015). "Identification and characterization of a serious multidrug resistant Stenotrophomonas maltophilia strain in China." *Biomed Res. Int.* 2015, p. 580240. DOI: 10.1155/2015/580240.

Ziegler-Heitbrock L, Ancuta P, Crowe S, et al. (2010). "Nomenclature of monocytes and dendritic cells in blood". *Blood* 116.16, e74–e80. DOI: 10.1182/blood-2010-02-258558.

Ziegler-Heitbrock L and Hofer TPJ (2013). "Toward a refined definition of monocyte subsets." *Front. Immunol.* 4.FRB, p. 23. DOI: 10.3389/fimmu.2013.00023.

Zimlichman E, Henderson D, Tamir O, et al. (2013). "Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system." *JAMA Intern. Med.* 173.22, pp. 2039–46. DOI: 10.1001/jamainternmed.2013.9763.

Zompi S, Montoya M, Pohl MO, Balmaseda A, and Harris E (2012). "Dominant cross-reactive B cell response during secondary acute dengue virus infection in humans." *PLoS Negl. Trop. Dis.* 6.3, e1568. DOI: 10.1371/journal.pntd.0001568.

COLOPHON

This document was typset using XeTeX, starting from a fork of Tiffany Tseng's `tufte-latex-mit` template,[1] which is itself an amalgamation of components of MIT's LaTeX thesis template[2] and the `tufte-latex` template.[3] It unabashedly imitates the style of Aaron Turon's immaculately typeset PhD dissertation;[4] I hope he can forgive me for emulating an impressively optimal solution.

Body text is set in Minion Pro; monospaced text uses `Bitstream Vera Sans Mono`. Latin Modern is used for the pittance of math herein. **Cochin** is used for chapter numbers, while *Palatino* stealthily supersedes Minion Pro for chapter titles.

Source code for the typesetting of this thesis is available online.[5] I hereby disclaim any responsibility for the amount of procrastination said code enables in the course of writing your own dissertation.

[1] https://github.com/ttseng/tufte-latex-mit

[2] http://web.mit.edu/thesis/tex/

[3] https://tufte-latex.github.io/tufte-latex/

[4] Turon (2013), "Understanding and Expressing Scalable Concurrency".

[5] https://github.com/powerpak/thesis