Building a parasite knowledgebase for comparative genomics and modeling

Maureen A. Carey[*]
Gregory L. Medlock[†]
Michal Stolarcyzk[‡]
Jennifer L. Guler[§]
Jason A. Papin[¶]

ABSTRACT

Eukaryotic single-celled parasites cause many neglected tropical diseases, such as malaria, African sleeping sickness, diarrheal disease, and leishmaniasis, with diverse clinical presentations and large global impacts. Research focused on these pathogens has been limited due to economic and experimental constraints. Accordingly, data collected in one eukaryotic pathogen are frequently extrapolated to infer knowledge about another parasite, across and within genera. Model *in vitro* (like representative apicomplexan parasite, *Toxoplasma gondii*) or *in vivo* systems (like mouse models of disease) are frequently used due to their enhanced experimental manipulability. However, characterization of functional differences among parasite species is limited to *post hoc* and single target studies, limiting the utility of this extrapolation approach. To address this challenge, we present a functional comparative analysis of 162 genomes representing 111 *Plasmodium*, *Toxoplasma*, *Cryptosporidium*, *Entamoeba*, *Trypanosoma*, *Leishmania*, and *Giardia* species, using primarily metabolic modeling given the frequency of metabolic drug targets. We generated metabolic reconstructions to serve as a knowledgebase for each parasite and to use for comparative functional genomics and identified differences in gene essentiality and pathway utilization to facilitate comparison of experimental findings. Using this knowledgebase, we can identify species-specific functions, interpret experimental results, and optimize selection of experimental systems for fastidious species.

INTRODUCTION

Malaria, African sleeping sickness, many diarrheal diseases, and leishmaniasis are all caused by eukaryotic single-celled parasites, result in over one million preventable deaths annually, and contribute to a significant reduction in disability-adjusted life years. This global health burden makes parasitic diseases a top priority of many economic development and health advocacy groups. However, effective prevention and treatment strategies are lacking. No vaccine exists for any of these diseases. Patients have limited treatment options because few drugs exist for many of these diseases, drug resistance is common, and many drugs have stage specificity. Thus, there is a pressing need for novel, effective therapeutics.

Beyond the economic constraints associated with antimicrobial development, antiparasitic drug development is technically challenging for three primary reasons: these parasites are eukaryotes, have complex life cycles, and are challenging to grow *in vitro*. To elaborate, unlike prokaryotic pathogens, these parasites share many targetable features with their eukaryotic host. To overcome the similarity between host and pathogen, strategies similar to the development of cancer therapeutics are necessary to minimize the negative effects on host. Enzyme kinetics can be leveraged such that the drug targets the pathogen's weak points while remaining below the lethal dose for host (Haanstra et al. 2017). Alternatively, selection of pharmacological treatment can synergize with the host immune response (*e.g.* Bogdan et al. (1991) and Kumaratilake et al. (1997)). For example, parasites must often survive high redox stress caused by host immune cells; a secondary redox stressor (*i.e.* drug) can synergize with this host response. Unique parasite features (*i.e.* signalling cascades (*e.g.* Zheng (2013)) or plastid organelles (*e.g.* Dahl et al. (2006)) can also be targeted, if identified.

Secondly, drug development itself is hampered by the parasites' complicated life cycles in one or more hosts. The malaria parasites infect multiple tissue types in host (primate, rodent, bird, or reptile) and vector (mosquito); all of these stages are required for complete development. These diverse environmental conditions are hypothesized to maintain redundancy in each parasite's genome, as genes and functions may only be utilized during some life stages; thus, drug development must focus on function, not merely presence, of putative essential genes.

Many of these organisms have no *in vitro* and *in vivo* culture systems, like *Plasmodium vivax* (human malaria) and *Cryptosporidium hominis* (diarrheal disease). Some parasite species have additional unique experimental challenges hindering drug development, like resisting genetic modification. *Plasmodium falciparum*, the most lethal human malaria parasite, was considered refractory to genetic modification until recently (Ghorbal et al. 2014; Lee and Fidock 2014) due to extremely low transfection efficiency; *E. histolytica*, a diarrheal pathogen, has also been refractory to efficient genetic manipulation. The genomes of *Leishmania*, which causes ulcers, develop significant aneuploidy when under selective pressure due to genomic flexibility (Downing et al. 2011; Sterkers et al. 2012).

Although many of these challenges can be circumvented with new technology, the use of clinical samples, and reductionist approaches, there are minimal historic data for reference. Without adequate profiling data (genome-wide essentiality, growth profiling in diverse environmental conditions, etc.), we do not have the knowledge to rationally identify novel drug targets. Untargeted and unbiased screens of chemical compounds for antiparasitic effects have proven useful (if the organism is culturable), but this approach makes predicting and understanding drug resistance and resistance mechanisms challenging. As a result, data collected in one organism are frequently extrapolated to infer knowledge about another parasite, across and within genera. Trypanosomes and *Toxoplasma* are frequently used as model organisms for other parasites due to their genetic and biochemical manipulatability. Mouse models of malaria and cryptosporidiosis are caused by different species within the same genera as their human analogs. However, the modest characterization of functional differences among parasite species limits the utility of this extrapolation approach. Computational approaches, such as comparative genomics and modeling, can address these challenges by facilitating rigorous comparisons of eukaryotic organisms to increase the utiliity of extrapolation-based knowledge transfer. Genome-scale metabolic modeling, for example, provides a framework to understand parasite genomes, highlight knowledge gaps, and generate high-confidence data-driven hypotheses. Metabolic models are built from genomic data and by inferring function to complete or connect metabolic pathways; these reconstructions are supplemented with functional genetic and biochemical studies, representing our best understanding of an organism's biochemistry and cellular biology. Here, we present a parasite knowledgebase, **Para**site **D**atabase including **G**enome-scale metabolic **M**odels (**Paradigm**), for this purpose. With Paradigm, we compare metabolic capacity, gene essentiality, and pathway utilization to better leverage experimentally tractable model systems for the study of eukaryotic parasites and antiparasitic drug development.

RESULTS

Comparative genomics

Comparative genomics in the field of eukaryotic pathogens and apicomplexan parasites has primarily been focused on the study of parasite surface proteins that interact with the host (**REFS**). Accordingly, we first explore an unbiased comparative genomics analysis using 162 publically available genome sequences from the EuPathDB databases (**Table ??**); each EuPathDB is a rough phylogenetic grouping containing only organisms from one genus or several closely-related genera. Sequence-based analyses investigating genetic similarity can be biased by AT content, genome alignment and assembly, and structural genomic variants, and eukaryotic parasites include organisms with *X*-*X*% AT content, *X*-*X* chromosomes or contigs, and potential significant structural variation (*e.g.* *X*-*X* size, presence of plastid genomes, *X*-*X* chromosomes; **Table ??**). Accordingly, we analyzed amino acid sequences by examining a conserved open reading frame across nearly all genomes (**Suppl. Fig. ??**). Lactate dehydrogenase sequence clusters by genera, but it is challenging to interpret meaningful similarities and differences across genera (**Suppl. Fig. ??**).

To explore functional genomic content, we reannotated all genomes using Diamond against OrthoMCL genes and compared annotated genes in each genome. Each genome has unique gene annotations but many are shared (examples in **Fig. ??**). For example, *Trichomonas vaginalis* G3 is the only complete genome on the *Trichomonas* database (TrichDB); this genome has the second most unique annotations by genome and TrichDB has the third most unique annotations by genome despite containing only one genome. Unsurprisingly, some of the larger genomes, including *Chromera velia* CCMP2878 (CryptoDB, 193.4 megabases), *Acanthamoeba castellani* Neff (AmoebaDB, 42 megabases), and *T. vaginalis* G3 (176.3 megabases) have the most unique annotations (**Fig. ??**). Annotation similarities can generate novel hypotheses about functional similarities. For example, the largest overlapping annotation group (CryptoDB and AmoebaDB) contains two types gut pathogens, the causes of cryptosporidiosis and amoebiasis, and these shared annotations may be consistent with mechanisms of gut pathogenesis (**Fig. ??**). Similarly, *Chromera velia* CCMP2878 and *Vitrella brassicaformis* CCMP3155 share a large number of annotations, likely representing the photosynthetic functions encoded by the two genomes (**Fig. ??**). However, there are unique gene annotations associated with EuPathDB database (**Fig. ??**), and it is unclear whether these differences arise from divergent metabolic functionality or incomplete genome annotation of these enzymes (**Table ??**).

Models as knowledgebases

To address this challenge, we generated a metabolic reconstruction for each species using a novel pipeline (**Fig. ??**). Genome-scale metabolic reconstructions are built from genomic data and by inferring function to complete or connect metabolic pathways; these reconstructions can be supplemented with functional genetic and biochemical studies (see **Methods**). In brief, our pipeline included generating a *de novo* reconstruction for each genome (**Fig. ??A**). We next generated a semi-curated reconstruction for a subset of organisms by transforming a well-curated reconstruction (example shown in **Fig. ??B** using iPfal18, curation in [Carey, Papin, and Guler (2017); **REF ANA's PAPER**] and in Supplemental Results) using genetic orthology (**Fig. ??C** and **D**). Draft and semi-curated reconstructions were then gapfilled to produce biomass or complete literature-derived biochemical requirements. Parasites were compared using draft reconstructions at the *de novo* reconstruction step (**Fig. ??**, **??**, **??**, **??**, **INCLUDE ALL FIGURES**) and using semi-curated reconstructions at the automated orthology-driven curation step (**Figure**), as well as using the predictions generated by the final gapfilled model (**Figure**).

Our *de novo* (draft) reconstructions contain only genetically supported information (**Fig. ??A**), and reconstruction size correlates with genome size (**Fig. ??A**). Unsurprisingly, the large genome of *Chromera velia* CCMP2878 (CryptoDB, 31,799 ORFs, 2,943 reactions) has the most unique reactions (92, **Fig. ??**). However, even small reconstructions contain unique reactions (**Fig. ??B**). In fact, all reconstructions contain at least one unique reaction (**Fig. ??B**), and small reconstructions do not have fewer unique reactions (**Fig. ??C**). A core set of reactions are contained in all 162 reconstructions (right side of

Fig. ??D), and a large set of reactions are shared by only a few models (left side of **Fig. ??D**). Reactions shared by all models include functions such as glycolysis. **HOW DOES GAPFILLING CHANGE THIS**

Niche-specific metabolic functions

We next compare network structure and the predictions generated by each model, as we compared genomic content and annotations (**Fig. ??**, **??**, and **??**). Network structures were minimally overlapping with 40 reactions shared by all reconstructions and 999 reactions in at least 50% of models. By comparing metabolic reactions in each reconstruction, we compare metabolic capacity of each species; two pairs, first *P. falciparum* 3D7 and IT and second *P. yoelii* yoelii YM and 17X, were most similar and *C. velia* CCMP2878 and *T. cruzi* CL-Brener were most different, in contrast with the genetic similarity (**Figure ??**). As expected, models generated from genomes in the same genus contain similar sets of reactions (**Figure ??**). However, phylogeny does not dictate model similarity, as models often cluster within environmental niche rather than phylogenetic grouping. Apicomplexan parasites cluster tightly within genus but not across genus (**Figure**). *Cryptosporidium* parasite cluster with *Giardia* and other gut pathogens rather than other Apicomplexa (**Figure**).

To explore each parasite's metabolic dependence on their host cells, we identified metabolites that could be imported via a genetically-supported transporter. We conducted a pairwise similarity between the set of metabolites that could be imported in each reconstructions. Following classical multidimensional scaling (or principal coordinates analysis, **Figure ??A**), we compared transporter topology between genera and parasite groups. Reconstructions from organisms in the same genera had similar transport ability (**Figure ??A**); additionally, reconstructions separate by some host cell types, like the organism's ability to divide extracellularly (**Figure ??B**) or in a host red blood cell (**Figure ??C**).

Next, we performed automated curation. All reconstructions were gapfilled to ensure the network could consume or produce metabolites identified in **Supplemental Table X**. This gapfilling step required *X-X* additional reactions per reconstruction. Next, all *Plasmodium* reconstructions were semi-curated using our automated curation pipeline and the curated reconstruction, iPfal18 of *Plasmodium falciparum* metabolism (**Figure ??**) and gapfilled to generate functional networks (*i.e.* networks that could product ATP and 'grow' as measured by the ability to produce biomass). Many modification were made to each *Plasmodium* reconstruction following semi-curation (**Figure ??C** and **Table ??**), greatly improving the genome-wide coverage of the reconstructions. Lastly, reconstructions were gapfilled to generate biomass (see **Biomass** in **Methods**), adding X-X reactions and X-X metabolites.

Predicting metabolic function

To evaluate these networks, we compare *in silico* predictions to experimental results, like gene essentiality and biochemical essentiality studies. These datasets were available for only *Plasmodium falciparum* and *berghei*, *Toxoplasma gondii*, and *Trypanosoma brucei* (**Table ??**) and implemented for *Plasmodium falciparum* and *berghei* (**Figure ??**). Reaction networks obtained via *de novo* model construction or only orthologous transformation are similar in accuracy for *P. berghei* (**Figure ??A**), but the set of correct gene essentiality predictions are different using the two different networks, motivating the integrated approach presented in **Figure ??**. The semi-curated reconstructions are larger in scope due to the addition of reactions associated with genes added via orthologous-transformation, and their gene essentiality accuracy is similar to or more accurate than their corresponding draft reconstruction (**Figure ??B**). Finally, with our semi-curated reconstructions, we compared metabolic networks to identify divergent metabolism and pathways in which model organisms were similar. By exploring where each network fails to predict essentiality, we generate targeted experimental hypotheses for exploring differences between species and improving genome annotation.

We tested accuracy of model predictions from the *de novo* reconstruction, the orthology-translated reconstruction, and the final semi-curated reconstruction for *P. bergehi* and compared these summary statistics to the prediction accuracy generated by our well-curated *iPfal17* (**Figure ??**). This comparison was used to motivate our approach over *de novo* reconstruction building as our pipeline generates a reconstruction more accurate than *de novo* reconstruction and comparable to a well-curated reconstruction.

DISCUSSION

Here, we presented 162 novel metabolic reconstructions for major human pathogens and closely-related species and a pipeline for generating high-quality reconstructions from genomes (**Figure ??**). These reconstructions represent the first genome-scale metabolic reconstructions for many of these organisms, making Paradigm the broadest biochemical database for eukaryotic parasites to date. Paradigm uses BiGG reaction and metabolite nomenclature (King et al. 2016), facilitating future work involving host-pathogen interaction modeling, and EuPathDB gene nomenclature (@ Aurrecoechea et al. 2017), consistent with field standards. Reproducible data integration approaches are used to curate each reconstruction; code and data formating are available in the **Supplemental Material**. Our draft reconstruction approach contains key features to generate comprehensive networks for eukaryotic cells, making it unique among existing automated network reconstruction pipelines. Our semi-curation approach leverages the curation conducted in manually curated reconstructions for closely-related organisms and genetic orthology, generating reconstructions that are more comprehensive than draft reconstructions. Both draft and semi-curated reconstructions can be used for comparative analyses, further curated by the modeling community, and applied to interrogate clinically and biologically relevant phenotypes.

Our approach has several key features tailored to eukaryotic pathogens. For example, discussion of biomass formulation is sorely lacking in many novel reconstruction papers and the assumptions used in formulated a biomass reaction for prokaryotes may not apply to eukaryotes. These assumptions are important as the objective function (like a biomass reaction) influences gapfilling and essentiality analyses. For example, in the first genome-scale metabolic model of any *Cryptosporidium* species, *C. hominis* (Vanee et al. 2010), 30 of 117 reactions involved in lipid synthesis were unsupported by genetic evidence. The selection of biomass precursor metabolites like lipid species impact these results; for example, the 30 gapfilled reactions might not be added if alternative or fewer lipid species were included in the biomass reaction. Thus, to address these biomass-induced biases, we used multiple objective functions (ATP or biomass synthesis) and performed each gapfilling query 10 times to add confidence to our gapfilled reactions. For our semi-curated reconstructions, we also generated biomass reactions at multiple different scales: a universal and a genus-level biomass. We gapfilled each model to each of these objective reactions and added confidence scores to gapfilled reactions, corresponding to the number of gapfill solutions in which the reaction was added. These confidence scores inform our interpretation of model predictions (*i.e.* predictions involving low-confidence gapfilled reactions are low-confidence predictions) and highlight reactions for future manual curation. While including all gapfilled reactions (as opposed to just one possible solution) is not standard within the field, previous work has highlighted the uncertainty in network structure that gapfilling introduces (Biggs and Papin 2017). Thus, we believe that this uncertainty should be presented for future users and our confidence scores are a novel way to summarize this uncertainty.

Similarly, compartmentalization can induce biases in a model's predictions, as demonstrated in Carey, Papin, and Guler (2017). Compartmentalization is particularly relevant for generating reconstructions for eukaryotic organisms and a weak step of automated reconstruction approaches. To our knowledge, no automated approach addresses compartmentalization and, thus, compartmentalization is always added manually added. Both our *de novo* reconstruction and orthology-driven approaches addresses this. Compartmentalization was incorporated into our *de novo* reconstruction pipeline and implemented for several genera (**Table ??**). Furthermore, we used a curated model to inform the compartmentalization of each semi-curated model; genes associated with compartmentalized reactions were mapped via orthology, assuming orthologous genes has comparable localization across species. This adds compartments unique to these organisms, like the apicoplast for apicomplexan parasites (*i.e. Plasmodium*).

However, our approach regarding compartmentalization yields one principle weakness; no curation regarding the removal of functionality is implemented. For example, because genetically supported reactions were added to all feasible compartments, this adds plausible hypothetical network functionality. If a gene-encoded enzyme maps to mitochondrial and cytoplasmic reactions in an organism that contains a mitochondria, both versions will be included, adding network redundancy that may not be biologically accurate. Alternatively, if an enzyme maps to a chloroplast reaction that is not included in the BiGG database in any other subcellular compartment, we moved the reaction to the cytosol. It is plausible that chloroplast reactions like this example are not catalyzed by the parasite. However, it is likely that parasite do have functionality not well summarized in this database, which contains no parasite reconstructions, but 6 mammalian, 5 other eukaryotic, 52 *E. coli*, and 12 other bacterial reconstructions (King et al. 2016). These modifications are encoded in our analytic pipeline for future reference (see code, linked here). Consequently, our reconstructions will require manual network curation especially regarding pruning of excess functionality. This is also a weakness of our orthology-driven curation approach, which adds function without function removal, and of many modeling construction and validation (*i.e.* metabolic tasks) approaches as it is difficult to validate lack of function.

Our approach generated more accurate and comprehensive models. To evaluate these networks, we compare *in silico* predictions to experimental results, like gene essentiality to targeted knockouts and genome-wide screens. Importantly, we tested accuracy of model predictions from the *de novo* reconstruction, the orthology-translated reconstruction, and the final semi-curated reconstruction for *P. bergehi* and compared these summary statistics to the prediction accuracy generated by our well-curated *iPfal17*. This comparison was used to motivate our approach over *de novo* reconstruction building as our pipeline generates a reconstruction more accurate than *de novo* reconstruction and comparable to a well-curated reconstruction (**Figure ??**). However, all of our models have imperfect accuracy. False positives, or when the model incorrectly predicts a gene or enzyme as essential, are a product of the model building process; these reconstructions are built to summarize *all* metabolic capabilities of the organism, not the specific stage-dependent phenotype of an organism in the experimental system. Thus, constraining a reconstruction with *in vitro* expression data will reduce the false positive rate. False negatives, or when the model incorrectly predicts a gene or enzyme as nonessential, highlight gaps in our understanding of the organism's metabolism. These can be pursued when manually curating individual networks.

Directly answering our motivating biological question, we compared metabolic networks to identify divergent or conserved metabolic pathways to better leverage model systems for drug development. Network structures were quite unique with only 0.19952% of all reactions in more than 50% of the reconstructions; network topology did however clustered by genus (**Figure ??**) and transport ability is associated with host environment (**Figure ??**).

Despite structural similarities, minor topological differences in networks confer key metabolic strengths or weaknesses.

Thus, we present a Paradigm, a framework for comparing and contrasting eukaryotic parasites and their metabolic function. We demonstrate the utility of this framework by identifing several novel findings, not readily apparent by genomic analysis alone. First, all parasite genomes encode unique metabolic functions, regardless of genome size, and parasites within the same genera tend to have similar network topology overall. Parasite may also have convergently evolved to their metabolic niche as enteric pathogens, for example, share metabolic functions and host cell type is associated with genetically-encoded transport ability from the extracellular environment to the parasite cytoplasm. Lastly, networks vary in the number and ratio of phosphate-using reaction they contain and the effect of this must be explored in inhibitor screens.

Paradigm provides a framework for organizing and interpreting our biochemical knowledge about eukaryotic parasites. This framework implements and builds on field-accepted standards for genome-scale metabolic modeling and the latest genome annotations in the parasitology field and can be implemented with other organisms, eukaryotic or prokaryotic. Paradigm, specifically, can be used broadly by the community and re-implemented iteratively to incorporate new genome sequences, novel datasets, and genome annotation updates. We call these networks 'semi-curated' to differentiate between the commonly used and referenced, uncurated 'draft' and well-curated network states. However, each reconstruction will require additional manual curation to maximize the utility and predictive accuracy. These reconstructions can be used to generate targeted experimental hypotheses for exploring differences between species and improving genome annotation by exploring differences between *in vitro* observations and *in silico* predictions. By applying this approach, we aim to develop a framework for identifying the best *in vitro* system or non-primate infection model of disease for drug development, and hypothesize that the best test system may vary by metabolic pathway for any one human pathogen.

SUPPLEMENTARY DATA

ACKNOWLEDGEMENTS

FUNDING

This work was supported by the National Institutes of Health [XXX to GLM and XXX to all other authors].

CONFLICT OF INTEREST

REFERENCES

TABLE AND FIGURES LEGENDS

METHODS

All code is available on GitHub, linked here. R (R Core Team 2017) and R packages tidyverse, ggdendro, seqinr, Biostrings, msa, reshape2, UpSetR, and ggdendro were used for analysis or visualization (Wickham 2017, 2012; Vries and Ripley 2013, 2013; Charif and Lobry 2007; Pages et al., n.d.; Bodenhofer et al. 2015; Gehlenborg 2017). Python 3.6.4, pandas, CobraPy 0.13.0 (Ebrahim et al. 2013), and Medusa (**CITE**) and code from CarveMe (Machado et al. 2018) and Memote (Lieven et al. 2018) were used for genome-scale metabolic modeling.

Genomic Analyses: Sequences were obtained from EuPathDB release 39 (Aurrecoechea et al. 2017). EuPathDB curates and compiles genome annotation for all genomes hosted by the database. We used open reading frames identified on EuPathDB and supplemented EuPathDB functional annotations with *de novo* Diamond annotations, described below. EuPathDB's OrthoMCL was used for mapping orthology between species. In brief, orthology was mapped within each EuPathDB database by the 'map by orthology' tool from the genome of each organism with a curated reconstruction to all other genomes within that database. The search protocol was 'new search > genes > taxonomy > organism [pick] > transform by orthology'. We mapped each organism's amino acid sequences using Diamond annotation (Buchfink, Xie, and Huson 2015) against proteins referenced in the BiGG databases (King et al. 2016) or against protein sequences obtained from OrthoMCL, part of EuPathDB that contains orthologous groups of parasite genes (Li, Stoeckert, and Roos 2003). Diamond is a similar approach to BLAST, with sensitive and fast performance on protein annotations (Buchfink, Xie, and Huson 2015).

Model Generation: We generated draft reconstructions by first annotating each organism's amino acid sequences, obtained from EuPathDB (Aurrecoechea et al. 2017), using Diamond annotation (Buchfink, Xie, and Huson 2015) against proteins referenced in the BiGG databases (King et al. 2016). We next mapped all functional annotations to reactions contained in the BiGG database (King et al. 2016) inspired by the approach conducted with the reconstruction pipeline CarveMe (Machado et al. 2018). Methods are included in the analytic code hosted on my GitHub page, linked here.

Unlike the CarveMe approach (Machado et al. 2018), we included all high-scoring reactions rather than maximizing the number of high-scoring hits while building a functional network. This conservative approach generates broadly inclusive but incomplete reconstructions (*i.e.* that are not able to produce biomass until gapfilled). This approach added redundant reactions from multiple different compartments (*e.g.* peroxisome, mitochondria, and cytosol) so all reaction versions other than the cytosolic version were pruned unless contained in a relevant compartment (**Table ??**); for genera not included in **Table ??**, only the cytosol and extracellular space were used. For example, if a *Plasmodium* reconstruction contained a reaction in the cytosol, mitochondria, and chloroplast, only the cytoplasmic and mitochondrial reaction versions would be kept. Following this step, a large percentage of each reconstruction's reactions remained in unsupported compartments because there was no analogous cytosolic reaction. Next, reactions only found in an unsupported compartment were moved to the extracellular space or cytosol; specifically, periplasmic metabolites were moved to the extracellular space and all internal subcompartment metabolites were moved to the cytosol. However, this step removed all reactions that summarized a transport reaction from the extracellular space to periplasm or from the cytosol to an unsupported organelle. Note, the extracellular compartment corresponds to the parasitophorous vacuole space contained within the host cell for intracellular parasites (*i.e. Plasmodium*, *Toxoplasma*, *Cryptosporidium*) and the host serum for extracellular parasites (*i.e. Trypanosoma*).

Manual Curation: We performed brief manual curation from literature sources, building on our curation conducted in Carey, Papin, and Guler (2017) and **REF ANA's PAPER**. **Table ??** contains all modifications resulting from our literature review; see code for implementation. Networks were manually curated with 8 modifications to improve our asexual blood-stage *Plasmodium falciparum* 3D7 reconstruction iPfal17 (Carey, Papin, and Guler 2017), generating iPfal18.

Additional manual curation was performed on lipid metabolism of the asexual blood-stage *P. falciparum* using the lipidomics study presented in Gulati et al. (2015) (**Supplemental Table X**) adding *XX-XX* reactions, *XX-X* metabolites, and *XX-X* genes. This curation removed aggregate reactions representing lipid metabolism and replaced them with individual reactions for individual lipid species, as supported by the lipidomics study. This model is available (**Supplemental Material**), but was not used for the analyses presented here as the metabolic demands for lipids in our biomass reaction are also aggregated. Inclusion of these reaction is appropriate for understanding lipid metabolism but would create random distributions of flux through the individual reactions that may distract from meaningful changes in flux distributions.

Automated orthology-driven curation: We developed a novel automated curation approach using orthologous transformation, similar to the approach taken by Abdel-Haleem et al. (2018). Our approach leverage the curation conducted in one organism for closely-related organisms. We applied this approach to all draft *Plasmodium* reconstructions using iPfal18 (**Figure ??**). We first mapped orthology of *P. falicparum* to each other *Plasmodium* species to build an orthology thesaurus (**Figure ??C**). We then added genes and associated reactions from iPfal18 if there was an orthologous gene in the target species' reconstruction (**Figure ??D**) resulting in a mean 42 genes added (SD = 10.38) and a mean 113 reactions added (SD = 4.14, **Table ??**). Notably, this approach facilitates the compartmentalization of these reconstructions, a function many automated pipelines fail to include. This is particularly important for parasite-specific compartments like the apicoplast, which is not included in any database.

Models were tested for thermodynamically-infeasible loops and energy-generating cycles; the approach outlined in (**???**) was used with minor modifications for eukaryotic cells. See code, linked here, for details.

Automated data-driven curation: Further automated curation of all reconstructions was performed by gapfilling for metabolites measured to be consumed in fluxomic or select media formulation studies. Detailed analysis is provided in our analytic code, on Github, linked here. Following an extensive literature review, we compiled data providing evidence for consumption or production of select metabolites (**Supplemental table- auxotrophies references**). Metabolites were defined as consumed by the parasite if (1) the metabolite was radiolabeled, added to media, incorporated into the parasite or converted by the parasite, and this was not seen to the same degree in uninfected host cells (2) the metabolite rescued inhibitor treatment of a metabolically upstream parasite enzyme, or (3) the metabolite is an essential media component for parasite culture. Metabolites were defined to be produced by the parasite if the metabolite was radiolabeled following growth in a media containing a radiolabeled precursor metabolite, and this was not seen to the same degree in uninfected host cells. First, import or excretion of these metabolites were added to the reconstruction. Next, the model objective was changed to an internal demand reaction for the metabolite or excretion reactions, respectively, and was gapfilled sequentially; this ensures import or synthesis of each of these measured metabolites. Reaction added due to gapfilling were given confidence scores.

Gapfilling: Gapfilling is an analytic process used to bridge or complete genetically-supported metabolic pathways to permit the network to fulfill metabolic functions, and was used to generate functional models. To increase the scope of a reconstruction (*i.e.* to add reactions), we perform gapfilling to fill in gaps in a pathways to ensure that the reconstruction can complete a particular task. This optimization problem adds reactions to allow the reconstruction to carry flux under given constraints. The implications of this approach are described in **Biomass Formulation**. We scored gapfilled reactions to summarize the confidence of reaction addition. In short, each gapfilled reaction has a score associated with it for each type of gapfilling performed. We gapfilled for three or four objective functions (see next section) for ten iterations each. Gapfilling confidence is based on how frequently a reaction is added in any of the gapfilling solutions and is noted as follows. For example, a reaction with the score 'OF3_1:1' appeared in 100% of solutions, but only one solution was generated, whereas a reaction with score 'OF3_3:2' as necessary in two out of three solutions. These scores are formatted as ObjectiveFunction_Y:X, with 'ObjectiveFunction' indicating which objective functions were used for gapfilling (note: this ranges from three to four for this study), 'X' indicating the number of times a reaction is added, and 'Y' indicating the number of iterations used to solve each gapfilling problem. Gapfilling was conducted following all steps involving compartmentalization, manual curation, or automated curation.

Objective functions: We use two classes of objective functions here to robustly evaluate model performance. First, we maximize ATP production. Second, we use biomasses reactions, including a species- and genus-level curated biomass reactions (if available) and a generic biomass reaction. To generate species-level biomass reactions, we used biomass reactions from existing genome-scale metabolic models as drafts (**Table ??**). Species-specific reactions were used for all other species in the genera as the genus-level biomass reaction. Our generic biomass contains metabolites from several curated reconstructions (**Table ??**) and thus contains metabolites from the *Plasmodium falicparum*, *Leishmania major*, and *Cryptosporidium hominis* species-specific biomasses with the stoichometry contained in the iPfal18 biomass reaction. Unfortunately, variability in reconstruction namespace (*i.e.* the database used for metabolite and reaction nameing conventions) make it difficult to access data compiled for some parasite reconstructions, like the *Toxoplasma* and *Plasmodium* reconstructions shown, as there are not always one-to-one mappings of variable across databases. This generic biomass was used to capture the most conservatively defined required biosynthetic capacity.

Model Comparison: Resultant networks were compared by euclidean distance of reaction presence or of transporter capability. The BiGG reaction database includes alternative stoichiometries for several reactions; these alternate forms of a single reaction were interpreted as identical reactions. Transporter capability was identified by the presence of a reaction in the reconstruction (prior to gapfilling) that transported a metabolite from the extracellular compartment to the intracellular compartment. Thus, only genetically-supported transporters were analyzed.

Model Performance Evaluation: Network accuracy was evaluated against defined metabolic tasks (**Table ??**) and gene essentiality data (**Table ??**) if availale. In brief, metabolic tasks were defined from the literature and implemented as outlined in Carey, Papin, and Guler (2017). Gene essentiality is simulated by performing single deletion studies in our models. All simulations were performed in an unconstrained model (all exchange reactions were permitted to carry flux, simulating a nutrient rich environment); of note, unconstrained models do not necessarily represent the *in vitro* or *in vivo* environment in which all experiments were conducted, merely the metabolic capacity an organism encodes. Gene deletions were simulated by removing the gene of interest from the model. This change results in the inhibition of flux through all reactions that require that gene to function. If the model could not produce biomass with these constraints, the gene was deemed essential. Knockout accuracy was defined as the sum of true positives (refractory to knockout or lethal genes) and true negatives (nonlethal genes) divided by the total number of predictions. Biochemical studies were also used for model validation (**TABLE**). As targeted metabolomics data were used for model generation (by gapfilling), we excluded these data from the evaluation data, leaving untargeted metabolomics and enzyme inhibitions studies for validation.

Abdel-Haleem, Alyaa M, Hooman Hefzi, Katsuhiko Mineta, Xin Gao, Takashi Gojobori, Bernhard O Palsson, Nathan E Lewis, and Neema Jamshidi. 2018. "Functional Interrogation of Plasmodium Genus Metabolism Identifies Species- and Stage-Specific Differences in Nutrient Essentiality and Drug Targeting." *PLoS Comput. Biol.* 14 (1). journals.plos.org: e1005895. https://doi.org/10.1371/journal.pcbi.1005895.

Aurrecoechea, Cristina, Ana Barreto, Evelina Y Basenko, John Brestelli, Brian P Brunk, Shon Cade, Kathryn Crouch, et al. 2017. "EuPathDB: The Eukaryotic Pathogen Genomics Database Resource." *Nucleic Acids Res.* 45 (D1): D581–D591.

Biggs, Matthew B, and Jason A Papin. 2017. "Managing Uncertainty in Metabolic Network Structure and Improving Predictions Using EnsembleFBA." *PLoS Comput. Biol.* 13 (3). Public Library of Science: e1005413. https://doi.org/10.1371/journal.pcbi.1005413.

Bodenhofer, Ulrich, Enrico Bonatesta, Christoph Horejš-Kainrath, and Sepp Hochreiter. 2015. "Msa: An R Package for Multiple Sequence Alignment." *Bioinformatics* 31 (24): 3997–9. https://doi.org/

10.1093/bioinformatics/btv494.

Bogdan, Christian, Steffen Stenger, Martin Röllinghoff, and Werner Solbach. 1991. "Cytokine Interactions in Experimental Cutaneous Leishmaniasis. Interleukin 4 Synergizes with Interferon-γ to Activate Murine Macrophages for Killing of Leishmania Major Amastigotes." *Eur. J. Immunol.* 21 (2). Wiley Online Library: 327–33.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nat. Methods* 12 (1). nature.com: 59–60. https://doi.org/10.1038/nmeth.3176.

Carey, Maureen A, Jason A Papin, and Jennifer L Guler. 2017. "Novel Plasmodium Falciparum Metabolic Network Reconstruction Identifies Shifts Associated with Clinical Antimalarial Resistance." *BMC Genomics* 18 (1): 543.

Charif, Delphine, and Jean R Lobry. 2007. "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis." In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*, edited by Ugo Bastolla, Markus Porto, H Eduardo Roman, and Michele Vendruscolo, 207–32. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-35306-5/_10.

Dahl, Erica L, Jennifer L Shock, Bhaskar R Shenai, Jiri Gut, Joseph L DeRisi, and Philip J Rosenthal. 2006. "Tetracyclines Specifically Target the Apicoplast of the Malaria Parasite Plasmodium Falciparum." *Antimicrob. Agents Chemother.* 50 (9): 3124–31. https://doi.org/10.1128/AAC.00394-06.

Downing, Tim, Hideo Imamura, Saskia Decuypere, Taane G Clark, Graham H Coombs, James A Cotton, James D Hilley, et al. 2011. "Whole Genome Sequencing of Multiple Leishmania Donovani Clinical Isolates Provides Insights into Population Structure and Mechanisms of Drug Resistance." *Genome Res.* 21 (12). genome.cshlp.org: 2143–56. https://doi.org/10.1101/gr.123430.111.

Ebrahim, Ali, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. 2013. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python." *BMC Syst. Biol.* 7 (August). bmc-systbiol.biomedcentral.com: 74. https://doi.org/10.1186/1752-0509-7-74.

Gehlenborg, Nils. 2017. "UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets."

Ghorbal, Mehdi, Molly Gorman, Cameron Ross Macpherson, Rafael Miyazawa Martins, Artur Scherf, and Jose-Juan Lopez-Rubio. 2014. "Genome Editing in the Human Malaria Parasite Plasmodium Falciparum Using the CRISPR-Cas9 System." *Nat. Biotechnol.* 32 (8). nature.com: 819–21. https://doi.org/10.1038/nbt.2925.

Gulati, Sonia, Eric H Ekland, Kelly V Ruggles, Robin B Chan, Bamini Jayabalasingham, Bowen Zhou, Pierre-Yves Mantel, et al. 2015. "Profiling the Essential Nature of Lipid Metabolism in Asexual Blood and Gametocyte Stages of Plasmodium Falciparum." *Cell Host Microbe* 18 (3). Elsevier: 371–81. https://doi.org/10.1016/j.chom.2015.08.003.

Haanstra, Jurgen R, Albert Gerding, Amalia M Dolga, Freek J H Sorgdrager, Manon Buist-Homan, François du Toit, Klaas Nico Faber, et al. 2017. "Targeting Pathogen Metabolism Without Collateral Damage to the Host." *Sci. Rep.* 7 (January): 40406. https://doi.org/10.1038/srep40406.

King, Zachary A, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. 2016. "BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models." *Nucleic Acids Res.* 44 (D1). academic.oup.com: D515–22. https://doi.org/10.1093/nar/gkv1049.

Kumaratilake, L M, A Ferrante, B S Robinson, T Jaeger, and A Poulos. 1997. "Enhancement of Neutrophil-Mediated Killing of Plasmodium Falciparum Asexual Blood Forms by Fatty Acids: Importance of Fatty Acid Structure." *Infect. Immun.* 65 (10). Am Soc Microbiol: 4152–7.

Lee, Marcus Cs, and David A Fidock. 2014. “CRISPR-mediated Genome Editing of Plasmodium Falciparum Malaria Parasites.” *Genome Med.* 6 (8). genomemedicine.biomedcentral: 63. https://doi.org/10.1186/s13073-014-0063-9.

Li, Li, Christian J Stoeckert Jr, and David S Roos. 2003. “OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.” *Genome Res.* 13 (9). genome.cshlp.org: 2178–89. https://doi.org/10.1101/gr.1224503.

Lieven, Christian, Moritz Emanuel Beber, Brett G Olivier, Frank T Bergmann, Parizad Babaei, Jennifer A Bartell, Lars M Blank, et al. 2018. “Memote: A Community-Driven Effort Towards a Standardized Genome-Scale Metabolic Model Test Suite.” *bioRxiv.* https://doi.org/10.1101/350991.

Machado, D, S Andrejev, M Tramontano, and K R Patil. 2018. “Fast Automated Reconstruction of Genome-Scale Metabolic Models for Microbial Species and Communities.” *bioRxiv.* biorxiv.org.

Pages, H, P Aboyoun, R Gentleman, and S DebRoy. n.d. “Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms. R Package. 2014.”

R Core Team. 2017. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for Statistical Computing.

Sterkers, Yvon, Laurence Lachaud, Nathalie Bourgeois, Lucien Crobu, Patrick Bastien, and Michel Pagès. 2012. “Novel Insights into Genome Plasticity in Eukaryotes: Mosaic Aneuploidy in Leishmania.” *Mol. Microbiol.* 86 (1). Wiley Online Library: 15–23. https://doi.org/10.1111/j.1365-2958.2012.08185.x.

Vanee, Niti, Seth B Roberts, Stephen S Fong, Patricio Manque, and Gregory A Buck. 2010. “A Genome-Scale Metabolic Model of Cryptosporidium Hominis.” *Chem. Biodivers.* 7 (5). Wiley Online Library: 1026–39.

Vries, A de, and B D Ripley. 2013. “Ggdendro: Tools for Extracting Dendrogram and Tree Diagram Plot Data for Use with Ggplot.”

Wickham, H. 2012. “Reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.”

Wickham, Hadley. 2017. “Tidyverse: Easily Install and Load’tidyverse’packages.”

Zheng, Weiping. 2013. “Sirtuins as Emerging Anti-Parasitic Targets.” *Eur. J. Med. Chem.* 59 (January). Elsevier: 132–40. https://doi.org/10.1016/j.ejmech.2012.11.014.