UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging computational approaches for the identification of therapeutic drug target candidates in *Plasmodium* parasites

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Karla Patricia Godinez Macias

Committee in charge:

Professor Elizabeth A. Winzeler, Chair
Professor Nathan E. Lewis, Co-Chair
Professor Fleur M. Ferguson
Professor Theresa Gaasterland
Professor Michael K. Gilson

2025

The Dissertation of Karla Patricia Godinez Macias is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

TABLE OF CONTENTS iv

LIST OF SUPPLEMENTAL FILES viii

LIST OF FIGURES ix

LIST OF TABLES x

ACKNOWLEDGEMENTS xi

VITA xv

ABSTRACT OF THE DISSERTATION xviii

INTRODUCTION 1

Chapter 1 CACTI: An in silico chemical analysis tool through the integration of chemogenomic data and clustering analysis 3

1.1 Chemogenomic database selection and accessing 4

1.2 Querying and standardization of chemogenomic databases using SMILES. 5

1.3 Chemical comparison through similarity calculations 6

1.4 Drug target identification 8

1.5 CACTI, a large-scale chemical compound analysis tool 9

1.6 Results 10

1.7 Case study: assessment of drug-target querying using the Pathogen Box dataset 11

1.7.1 Module 1 identifies 4,315 synonyms for the Pathogen Box compounds 12

1.7.2 Module 1 identifies new citations, datasets and patents for Pathogen Box molecules 13

1.7.3 Module 2 identifies 12,383 new closely related compounds in the public domain that can be used for SAR by inventory 15

1.7.4 Module 3 for new target discovery hypotheses —guilt by association approaches 16

1.8 Challenges and limitations 18

1.9 Acknowledgements 21

APPENDIX 26

Supplemental Figure 1.1 Modules for drug target prediction 26

Supplemental Figure 1.2 PubChem duplicated analogs fingerprint similarity 27

Chapter 2 Systematic in vitro evolution in *Plasmodium falciparum* reveals key determinants of drug resistance 28

2.1 Introduction 29

2.2 Results 30

2.2.1 Compound-selected clones contain few mutations 30

2.2.2 Compound-selected mutations differ from passenger mutations 32

2.2.3 Copy number variants frequently drive compound resistance 33

2.2.4 Overrepresented genes affect compound resistance and culture adaptation 35

2.2.5 Network analysis reveals co-occurring mutations 36

2.2.6 Compound-selected mutations differ from naturally occurring variants 37

2.3 Discussion 38

2.4 Materials and methods 40

2.4.1 Data acquisition 40

2.4.2 In vitro evolution of compound-resistant parasites 40

2.4.3 Whole-genome sequencing analysis and annotation of variants 42

2.4.4 Gene network construction 43

2.5 Acknowledgements 44

APPENDIX 49

Supplemental Figure 2.1 Indels and SNV detection workflow. 49

Supplemental Figure 2.2 CNV detection workflow. 50

Chapter 3 Revisiting the *Plasmodium falciparum* druggable genome using predicted structures and data mining 51

3.1 Introduction 52

3.2 Results 53

3.2.1 Defining 1,660 *P. falciparum* genes with evidence of small molecule binding ... 53

3.2.2 Defining 1,929 *P. falciparum* genes with evidence of blood-stage essentiality ... 55

3.2.3 867 *P. falciparum* proteins have evidence of binding and blood-stage essentiality 57

3.2.4 Building an annotation resource using scientific evidence to prioritize candidate targets 58

3.2.5 Scoring 540 novel candidate targets with strong evidence of essentiality 60

3.2.6 Secondary scoring of 67 high-ranking candidate targets 62

3.2.7 In-depth consideration of 27 prioritized candidates reveals targets poised for drug discovery 63

3.3 Discussion 63

3.4 Methods 66

3.4.1 Data acquisition 66

3.4.2 Mapping genes to associated literature publications 66

3.4.3 Determining candidate proteins with evidence of small molecule binding 67

3.4.4 Identification of human orthologs 68

3.4.5 Definition of hypervariable and core genomic regions 69

3.4.6 Categorization of gene essentiality evidence 69

3.4.7 Assessment of gene expression by lifecycle stage 70

3.4.8 Candidate target scoring rubric 71

3.5 Acknowledgements 72

APPENDIX 79

Supplemental Figure 3.1 Essentiality categorization determination scheme for 5,318 *P. falciparum* protein-coding genes for target identification 79

Supplemental Figure 3.2 Data collected for full genome 80

Chapter 4 Dissecting imidazolopiperazine and naphthoquinone mechanisms of action in *Plasmodium* asexual malaria parasites.....81

4.1 Introduction.....82

4.2 Results.....84

4.2.1 Examining drug treatment during *Plasmodium* falciparum asexual blood-stage malaria infection.....84

4.2.2 *Plasmodium* asexual lifecycle patterns are different relative to after drug treatment.....85

4.2.3 Differential expression shows distinct drug response profiles dependent on the chemotype.....87

4.2.4 Finding *Plasmodium* falciparum orthologs between Dd2 and 3D7 strains for differentially expressed genes.....89

4.2.5 Pathway analysis of differentially expressed genes show hints about the regulatory mechanisms of atovaquone and gnf179.....91

4.3 Discussion.....92

4.4 Methods.....94

4.4.1 *Plasmodium* Dd2 parasites culture and sequencing.....94

4.4.2 Bioinformatics workflow.....95

4.4.3 Cell marker identification.....96

4.4.4 Differential gene expression.....97

4.4.5 Finding 3D7 orthologs.....97

4.4.6 Pathway enrichment analysis.....98

4.5 Acknowledgements.....98

APPENDIX.....107

Supplemental Figure 4.1 Clustering for artemisinin-, gnf179-, and atovaquone-treated cells versus non-treated cells.....107

Chapter 5 Conclusions.....108

REFERENCES.....113

LIST OF SUPPLEMENTAL FILES

Godinez-Macias_1-1_Pathogen-Box-references.xlsx – Manual curation and literature search for Pathogen Box from a comprehensive *in silico* literature search

Godinez-Macias_1-2_Pathogen-Box-analogs.xlsx – Close analog found in the public domain for Pathogen Box set.

Godinez-Macias_1-3_Pathogen-Box-clusters.xlsx – Chemical clustering for Pathogen Box compounds and their close analogs against a collection of known drug-target pairs.

Godinez-Macias_2-1_Sequenced-clones.xlsx – List of in vitro compound-selected clones metadata and sequencing statistics

Godinez-Macias_2-2_Clones-SNVs-Indels.xlsx – List of high-quality single nucleotide variants (SNVs) and indels in compound-selected clones

Godinez-Macias_2-3_Clones-CNVs.xlsx – List of copy number variants (CNVs) in compound-selected clones

Godinez-Macias_2-4_Compound-target-pairs.xlsx – Compounds classified based on target or resistance gene evidence, if applicable

Godinez-Macias_3-1_Pf3D7-binding-essentiality.xlsx – Small molecule binding and blood-stage essentiality evidence for all *P. falciparum* 3D7 protein-coding genes

Godinez-Macias_3-2_BindingDB-orthologs.xlsx – *P. falciparum* ortholog to UniProt protein IDs extracted from BindingDB having validated ligands.

Godinez-Macias_3-3_Target-candidates.xlsx – List of 867 *P. falciparum* candidate targets with evidence of small molecule binding and blood-stage essentiality.

Godinez-Macias_3-4_Candidates-ranking.xlsx – Ranking scores for 540 selected druggable candidate targets.

Godinez-Macias_4-1_DEG-analysis.xlsx – List of *P. falciparum* Dd2 list that were differentially expressed after treatment with artemisinin (n = 3,172), gnf179 (n = 1,488) and atovaquone (n = 2,239), with respect to control parasites.

Godinez-Macias_4-2_Enriched-pathways.xlsx – List of up- and down-regulated *Plasmodium* enriched pathways (Metascape) for ±1 fold-change when comparing drug-treated versus control Dd2 parasites.

LIST OF FIGURES

Figure 1.1 CACTI pipeline workflow 22

Figure 1.2 Identification of synonyms, references, and close analogs for 374 Pathogen Box compounds 23

Figure 1.3 Chemical clustering and target prediction 24

Figure 2.1 Compound-selected single nucleotide variants (SNVs) and indels have different characteristics compared to naturally occurring variants 45

Figure 2.2 Copy number variants (CNVs) are a frequent driver of antimalarial compound resistance 46

Figure 2.3 Enriched genes mutated in compound resistance selections 47

Figure 3.1 Identification of *P. falciparum* protein-coding genes that are potentially druggable (*n* = 1,660) and genes that have evidence of blood-stage essentiality (*n* = 2,992) 73

Figure 3.2 Characteristics of 867 candidate targets compared to 4,451 non-candidate genes 75

Figure 3.3 Rubric-based scoring of 540 candidate targets with strong evidence of essentiality 76

Figure 4.1 Experimental design and processing of *P. falciparum* asexual parasites 99

Figure 4.2 Transcriptional profiling of asexual *Plasmodium* parasites relative to after drug treatment 101

Figure 4.3 Enriched pathways relative to after drug treatment 103

LIST OF TABLES

Table 1.1 Number of records per database. 25
Table 2.1 Compound library used to test blood-stage parasites resistance. 48
Table 3.1 List of 67 "druggable target" candidates from the entire *P. falciparum* genome. 77
Table 4.1 Gene markers used for asexual stage cell type identification. 105
Table 4.2 Gene differences between schizont clusters in artemisinin and atovaquone treatment, and trophozoite clusters in artemisinin treatment. 106

ACKNOWLEDGEMENTS

I would like to thank my advisor, Elizabeth Winzeler, for her invaluable guidance, advice, and growth opportunities throughout my studies that shaped me into a better scientist. With her support, I collaborated in high impact projects and participated at invigorating national and international conferences. I also thank her for the understanding and support during difficult personal times. I would also like to thank my co-advisor, Nate Lewis, for his constant moral support and advice. I am thankful for the feedback during and after examinations and coursework, that helped in my scientific development and presentation skills. I also thank my doctoral committee members for their time, feedback, and their support throughout my studies. I am grateful to Terry Gaasterland, for her support on and off academic topics. It was a great pleasure TA'ing for you; your enthusiasm for teaching and dedication to student are truly inspiring. I also thank Mike Gilson, for his eagerness to collaborate and his support during the process. BindingDB has been an instrumental database for my research, and I am grateful for him and his team for building and maintaining the site. I also thank Fleur Ferguson for her valuable feedback throughout examinations. I would like to thank each of them for their time and support while completing my studies and research.

This work would not be possible without the support from incredible scientists and collaborators from the Malaria Drug Accelerator Consortium (MalDA). I am grateful for the support of Marcus Lee, F. Javier Gamo, David Fidock, Dyann Wirth, Amanda Lukens, Daniel Goldberg, Jacquin Niles, Ian Gilbert, Brice Campo, and all MalDA lead researchers and members. I am special grateful to have worked more closely with Lauren Coulson, John Okombo, Krittikorn Kumpornsin, and Stephen Patterson. Funding for my work was provided by the Gates Foundation, assisted by Gang Liu and Lincoln Wallis, whom I also thank.

I am very thankful to current, and former, Winzeler lab members. I want to give special thanks to Sabine Ottilie, who played a significant role as a mentor and friend. I thank her for getting me up to speed when I joined the lab and for her guidance throughout. I also want to thank Jan Economy, for helping me navigate administrative issues and answering all my questions. I also want to thank my fellow graduate colleagues Madeline Luth, Krypton Carolino, Daisy Chen, Priyan Kapoor, Fu-Hsuan Ko and Jason Hsiao for their feedback during group meetings. It has been a pleasure working with Frances Rocamora, Nimisha Mittal, Tuo Yang, Juan C. Jado, Justin Fong, Mariana Laureano, Tiantian Jiang and Jair Siqueira. Especially, I want to thank my friend Jaeson Calla, who was not only an amazing coworker and my go-to for scientific questions, but also for your friendship, the many lunch and coffee breaks, and endless laughs.

I would also like to thank my friend, Arjana Begzati, with whom I completed coursework, projects, and research papers. Her friendship outside work has helped me navigate graduate life more pleasantly. To my friends Marcela Johnson and Kabita Baral, for being there and sharing advice when I needed it. I want to thank my friend and former mentor, Jon Mohl, for training and helping me with many coding and bioinformatics questions. I also want to thank my friends Anais, Sofia, Lorena, Dasya, Eduardo, Ernesto and Raúl for their endless support and companionship.

I want to thank my family; they are my rock. To my parents, Paty and Victor, for always supporting me and being a role model; their care, values, sacrifices and love has shaped me into the person I am. To Fer, my sister, for pushing me to be a better version of myself. My grandparents, Silvia, Lulú and Carlos Macías, for sharing my accomplishments and loving me. I am grateful for my second family, Lulú, Phillip, Naty and Steven Ortega. You welcomed me with open arms when I moved abroad and made me feel like part of your family; my accomplishments are thanks to your support, and I am eternally grateful for that. I feel lucky to have met Everardo.

You are an important pilar in my life, cheering and supporting me no matter what, celebrating my accomplishments, and making me feel loved. I am so grateful for you and all you do for me. I also want to thank Charly, the Favila's and Navarro's, whom I have shared so many memories, weekend get-togethers and their encouraging words keep me motivated to keep going.

Chapter 1 includes a partial reprint of the material as it appears in Journal of Cheminformatics, 2024. Karla P. Godinez-Macias & Elizabeth A. Winzeler, "CACTI: an in silico chemical analysis tool through the integration of chemogenomic data and clustering analysis". The dissertation author was the primary author of this paper.

Chapter 2 includes a partial, modified reprint of the material as it appears in Science, 2024. Madeline R. Luth, Karla P. Godinez-Macias, Daisy Chen, John Okombo, Vandana Thathy, Xiu Cheng, Sindhu Daggupati, Heledd Davies, Satish K. Dhingra, Jan M. Economy, Rebecca C.S. Edgar, Maria G. Gomez-Lorenzo, Eva S. Istvan, Juan Carlos Jado, Gregory M. LaMonte, Bruno Melillo, Sachel Mok, Sunil K. Narwal, Tolla Ndiaye, Sabine Ottilie, Sara Palomo Diaz, Heekuk Park, Stella Peña, Frances Rocamora, Tomoyo Sakata-Kato, Jennifer L. Small-Saunders, Robert L. Summers, Patrick K. Tumwebaze, Manu Vanaerschot, Guoqin Xia, Tomas Yeo, Ashley You, Francisco-Javier Gamo, Daniel E. Goldberg, Marcus C.S. Lee, Case W. McNamara, Daouda Ndiaye, Philip J. Rosenthal, Stuart L. Schreiber, Gloria Serra, Jair Lage De Siqueira-Neto, Tina S. Skinner-Adams, Anne-Catrin Uhlemann, Nobutaka Kato, Amanda K. Lukens, Dyann F. Wirth, David A. Fidock, Elizabeth A. Winzeler, "Systematic in vitro evolution in *Plasmodium falciparum* reveals key determinants of drug resistance". The dissertation author was a co-first author of this paper.

Chapter 3 includes a partial, modified reprint of the material as it appears in NPJ Drug Discovery, 2025. Karla P. Godinez-Macias, Daisy Chen, J. Lincoln Wallis, Miles G. Siegel, Anna

Adam, Selina Bopp, Krypton Carolino, Lauren B. Coulson, Greg Durst, Vandana Thathy, Lisl Esherick, Madeline A. Farringer, Erika L. Flannery, Barbara Forte, Tiqing Liu, Luma Godoy Magalhaes, Anil K. Gupta, Eva S. Istvan, Tiantian Jiang, Krittikorn Kumpornsin, Karen Lobb, Kyle J. McLean, Igor M. R. Moura, John Okombo, N. Connor Payne, Andrew Plater, Srinivasa P. S. Rao, Jair L. Siqueira-Neto, Bente A. Somsen, Robert L. Summers, Rumin Zhang, Michael K. Gilson, Francisco-Javier Gamo, Brice Campo, Beatriz Baragaña, James Duffy, Ian H. Gilbert, Amanda K. Lukens, Koen J. Dechering, Jacquin C. Niles, Case W. McNamara, Xiu Cheng, Lyn-Marie Birkholtz, Alfred W. Bronkhorst, David A. Fidock, Dyann F. Wirth, Daniel E. Goldberg, Marcus C.S. Lee, Elizabeth A. Winzeler, "Revisiting the *Plasmodium falciparum* druggable genome using predicted structures and data mining". The dissertation author was a co-first author of this paper.

Chapter 4 contains unpublished material co-authored with Elizabeth A. Winzeler. The dissertation author was the primary author of this chapter.

VITA

2016 Bachelor of Science in Computer Science, Autonomous University of Baja California, Ensenada, Mexico

2016 Intern at Center for Scientific Research and Higher Education at Ensenada (CICESE), Ensenada, Mexico

2018 Intern at European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

2019 Master of Science in Bioinformatics, University of Texas at El Paso

2020 Research Assistant, University of California San Diego

2025 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Probst, A., Paton, D.G., Appetecchia, F., Bopp, S., Adams, K.L., Rinvee, T.A., Pou, S., Winter, R., Du, E.W., Yahiya, S., Vidoudez, C., Singh, N., Rodrigues, J., Castañeda-Casado, P., Tammaro, C., Chen, D., **Godinez-Macias, K.P.**, Jaramillo, J.L., et al. (2025). In vivo screen of *Plasmodium* targets for mosquito-based malaria control. Nature, https://doi.org/10.1038/s41586-025-09039-2.

Begzati, A., **Godinez-Macias, K. P.**, Long, T., et al. (2025). Plasma Lipid Metabolites, Clinical Glycemic Predictors, and Incident Type 2 Diabetes. Diabetes care, 48(**3**), 473–480.

Godinez-Macias, K. P., Chen, D., Wallis, J. L., et al. (2025). Revisiting the *Plasmodium falciparum* druggable genome using predicted structures and data mining. NPJ Drug Discovery 2(**1**):3.

Luth, M. R., **Godinez-Macias, K. P.**, Chen D., et al. (2024). Systematic in vitro evolution in *Plasmodium falciparum* reveals key determinants of drug resistance. Science. 386(**6725**):eadk9893.

Godinez-Macias, K. P., & Winzeler, E. A. (2024). CACTI: an in silico chemical analysis tool through the integration of chemogenomic data and clustering analysis. Journal of cheminformatics, 16(**1**), 84.

Jiang, T., **Godinez-Macias, K. P.**, et al. (2024). Identification of fungal natural products with potent inhibition in *Toxoplasma gondii*. Microbiology spectrum, 12(**4**), e0414223.

Mittal, N., Davis, C., McLean, P., Calla, J., **Godinez-Macias, K. P.**, et al. (2023). Human nuclear hormone receptor activity contributes to malaria parasite liver stage development. Cell chemical biology, 30(**5**), 486–498.e7.

Armstrong, J. F., Campo, B., Alexander, S. P. H., Arendse, L. B., Cheng, X., Davenport, A. P., Faccenda, E., Fidock, D. A., **Godinez-Macias, K. P.**, et al. (2023). Advances in malaria pharmacology and the online guide to MALARIA PHARMACOLOGY: IUPHAR review 38. British journal of pharmacology, 180(**15**), 1899–1929.

Ottilie, S., Luth, M. R., Hellemann, E., Goldgof, G. M., Vigil, E., Kumar, P., Cheung, A. L., Song, M., **Godinez-Macias, K. P.**, et al. (2022). Adaptive laboratory evolution in *S. cerevisiae* highlights role of transcription factors in fungal xenobiotic resistance. Communications biology, 5(**1**), 128.

Yang, T., Ottilie, S., Istvan, E. S., **Godinez-Macias, K. P.**, et al. (2021). MalDA, accelerating malaria drug discovery. Trends in parasitology, 37(**6**), 493–507.

Kellman, B. P., Zhang, Y., Logomasini, E., Meinhardt, E., **Godinez-Macias, K. P.**, et al. (2020). A consensus-based and readable extension of Linear Code for Reaction Rules (LiCoRR). Beilstein journal of organic chemistry, 16, 2645–2662.

Abraham, M., Gagaring, K., Martino, M. L., Vanaerschot, M., Plouffe, D. M., Calla, J., **Godinez-Macias, K. P.**, et al. (2020). Probing the open global health chemical diversity library for multistage-active starting points for next-generation antimalarials. ACS infectious diseases, 6(4), 613–628.

FIELD OF STUDY

Major Field: Bioinformatics and Systems Biology
Professor Elizabeth A. Winzeler

ABSTRACT OF THE DISSERTATION

Leveraging computational approaches for the identification of therapeutic drug target candidates in *Plasmodium* parasites

by

Karla Patricia Godinez Macias

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2025

Professor Elizabeth A. Winzeler, Chair
Professor Nathan E. Lewis, Co-Chair

Genomic databases help in the understanding of how genetic changes alter a phenotype and herein are linked to a disease. Similar resources are needed to study small molecules; they can alter a target gene's normal activity (e.g., drugs, ligands, inhibitors). But, contrary to genomic databases, chemogenomic databases are limited and less annotated. Furthermore, molecule

identifiers are not standardized and there are variations in naming depending on the database. This poses a challenge when trying to identify how a drug interacts with its biological target and whether it results in the desired phenotypic response. Knowledge on drugs' mechanism of action (MoA) or biological target helps translating those candidates into therapeutic treatments. In this dissertation, we use in silico methods to support the translation of genomic data into effective new drugs, by analyzing chemical libraries for early identification of research efforts that uses previously used compounds, close analogs or existing literature; and drug targets/MoA predictions. We find that by chemical querying and chemical comparisons we could provide target hypothesis for scaffold families and, paired with validation studies (e.g., in vitro evolution), it results in faster and more efficient drug development.

In Chapter 1, we describe CACTI, a tool that explores the chemical and biological space pragmatically, to provide an understanding of chemical structures and their impact on a biological system. This sets the basis for querying, comparing, clustering, and predicting potential targets for chemical compounds. In Chapter 2, we demonstrate that small molecules could have liabilities. Through the analysis of 724 *Plasmodium falciparum* genomes resistant to one of 118 small molecules, we show commonalities between resistant clones setting the foundation for machine learning predictive approaches. Chapter 3 describes the use of data mining and protein structure prediction to identify novel antimalarial drug targets, supporting the development of drugs that differ in MoA and resistance liabilities. Lastly, in Chapter 4, we explore the use of single-cell sequencing to provide clues about the mechanism for uncharacterized chemotypes. Contrasting transcriptomic profiles of *P. falciparum* parasites and drug-treated parasites, we found upregulation of biological pathways for atovaquone and artemisinin.

INTRODUCTION

Understanding how a drug interacts with its biological target and whether its inhibition results in the desired phenotypic response is a critical step in drug discovery[1-3]. One area where this is particularly true is for infectious diseases where many of the hits that are advanced for drug discovery come from phenotypic, organismal screens as described in a recent review[4]. For malaria parasites, tens of thousands of compounds with parasite killing activity have been placed in the public domain. This has led to medicinal chemistry programs and molecules entering clinical trials[4, 5]. While understanding the mechanism of action or knowing the target is not strictly essential for a compound series to advance, knowing the target and ideally having a crystal structure for the target can make subsequent medicinal chemistry optimization much more efficient: fewer compounds need to be made and evaluated, and biochemical assays are typically more robust and lower cost. Furthermore, screening and assay costs can be many orders of magnitude lower for a biochemical target relative to whole organism work.

Because of the huge-cost savings that can come from knowing a target, substantial effort is often invested into target discovery. Biochemical and genetic approaches (e.g. in vitro evolution) are two examples of methods used for target identification. However, these tend to be lengthy and are resource-consuming[4, 5]. For malaria parasites, in vitro evolution has been a successful method but can often take six months or more[4]. It thus makes sense to accomplish as much computationally as possible. For example, molecular docking is commonly pursued to predict small molecule ligand binding to key therapeutic proteins prior to biochemical or phenotypic assays, and though it tends to have a high false positive rate[6], drug targets for some antimalarial inhibitors have been identified through this approach[7-9]. Similarly, virtual screening is a commonly applied. Singh et

al.[10] describes how this approach can drastically reduce the chemical space to provide a focused library for subsequent biochemical high-throughput screening (HTS), although it is most reliable when knowledge of compound activity is available and libraries against a target of interest are accessible. Antimalarial compounds have also been elucidated through virtual screening[11-13]. Comparable in silico approaches also exist for target identification. For example, the tool TargetHunter by Wang et al.[14], is a web-based prediction approach that incorporates analog bioactivity data from ChEMBL[15]. Likewise, Chemmine, is an online resource by Backman et al.[16] that predicts targets based on similar records in PubChem database[17]. More recently, Jimenes-Vargas and colleagues[18] showed the power of artificial intelligence in drug discovery. Here, authors predict compound-target interactions based on different molecular description calculations and machine learning algorithms using ChEMBL records.

A constraint for these resources is the limit of one molecule per search and one chemogenomic database. Modern high-throughput screens may involve millions of compounds and thousands of primary hits. Having high-throughput methods to rapidly assess many hits can allow prioritization of compounds for resupply and resynthesis, which is particularly important for academic researchers that may have limited access to compound management groups or limited medicinal chemistry capacity. For example, in malaria parasites, many dihydrofolate reductase (DHFR) inhibitors have been compromised by widespread resistance[19]. It thus makes sense to identify phenotypic hits acting against DHFR prior to compound resupply. In fact, in silico approaches have proven to be a valuable in the analysis of large libraries, especially when various resources are utilized together to reduce the number of compound resupply and experimental testing needed[20, 21].

Chapter 1

CACTI: An in silico chemical analysis tool through the integration of chemogenomic data and clustering analysis

To address the need of an automated multi-compound analysis tool, particularly for neglected diseases, we constructed a pipeline named CACTI (Chemical Analysis and Target Identification), to provide comprehensive searches in chemogenomic databases and provide biological targets clues with single or bulk queries, integrating data not only from major databases such as ChEMBL and PubChem, but also from ligand-based databases (e.g. BindingDB[22]), as well as scientific and patent evidence (e.g. PubMed[23] and SureChEMBL[24]). Furthermore, to account for the difference in molecule identifiers across databases, we implemented a cross-reference method to map a given identifier based on chemical similarity scores and other known identifiers, or synonyms, in the expanded search. With this process, we provide a comprehensive large-scale study report incorporating all available identifiers for each small molecule, its close analogs, and available bioactivity data and/or mechanism of action (MoA) if known.

1.1 Chemogenomic database selection and accessing

A common first step to evaluate a small molecule as a potential therapeutic, includes the exploration of chemical and biological space to identify known information and potential target. In contrast to most available tools that focus on one database for target-predictions[14, 25-27] we explored multiple chemogenomic databases as knowledge source. The solely criterion for database selection was the availability of REST API, a data request protocol for query and transfer. For chemical- and experimental-based information sources, we selected the well-established ChEMBL and PubChem databases. BindingDB was selected for protein-ligand evidence and, EMBL-EBI and PubMed as sources for comprehensive searches (**Table 1.1**). Although each bears unique data, we observed some overlap in data content, especially for literature evidence. However, it is important to note the indexing type and database specification prior to integration, as not every

record is curated, and query fields could have different meaning. For example, experimental data in PubChem could be from biological assays (dose-response, phenotypic, or target-based) or biochemical assay (profiling or binding), while BindingDB refers to binding affinities assays solely.

After database selection, we created a custom function to access and retrieve data from each source. Following the corresponding REST API web services, we constructed an html query link with the database domain + web service architecture + data of interest + parameters. Specifically, we used chembl_id_lookup, mechanism, document, compound_record, molecule/synonym and similarity data functions from ChEMBL. For PubChem we used compound's data SMILES, name, CID, PubMedID, PatentID, fastsimilarity_2d, synonyms and MolecularWeight functions. PubMed records were found using the PMID and PMC xref from PubChem, and europepmc/search function from EMBL-EBI server. The command, GetTargetByCompound, was used to access BindingDB database. For exact URL query construction please refer to the database web-services[22-24, 28, 29].

1.2 Querying and standardization of chemogenomic databases using SMILES.

We reasoned that, to investigate a set of query compounds, the first step was to explore hidden patterns among the various chemogenomic databases containing large datasets of small molecules and their bioactivity. However, these patterns are difficult to reveal when searching across databases, mainly due to the lack of indexing standardization and the existence of multiple equivalent representations of a chemical structure (SMILES[30]). For example, ethanol can be encoded with SMILES OCC, CCO and C(O)C. Furthermore, in ChEMBL the SMILES CCO corresponds to CHEMBL545 but C(O)C is not a valid query, and vice versa the notation C(O)C

corresponds to CID702 in PubChem while CCO is not valid. Nevertheless, to test the hypothesis that using SMILES as identifier is sufficient for mapping the chemical space, we used the provided query SMILES as first input (**Supplemental Figure 1.1A**). To reveal the index identity of each query, we used the custom functions to cross-reference across databases using 100% similarity match, and further confirmed their exact identity by comparing Morgan fingerprints as described in our **methods**.

Due to submission discrepancies across repositories, query identifiers were combined in a "synonym" (common names) column and filtered if they were 1) numerical with no indication of external source origin, 2) IUPAC name from an unreliable source, or 3) duplicated when converted to upper case and removal of special characters such as ":" or "–". The remaining synonyms were used to retrieve scientific evidence, patent evidence and additional useful information associated with the compound of interest. We reasoned that expanding the search to include these identifiers would provide more exhaustive research on the query compounds, improving the target-prediction steps when using the additional pieces of evidence. With this filtering, invalid or duplicated query records were removed and bioactivity data, naming synonyms, scholarly evidence, and chemical information across selected chemogenomic databases was integrated.

1.3 Chemical comparison through similarity calculations

In addition to mining with SMILES and common names, we expanded the search to include closely related analogs (**Supplemental Figure 1.1B**). We first used RDKIT v.2024.03.1[31] to convert the query SMILES to a canonical form, generating a unique notation to be queried and compared with equal features. The standardized format (canonical) was used to identify analogs by transforming them to a binary representation called fingerprints (Morgan fingerprints), which

allows chemical similarity calculations. These similarities were computed using the Tanimoto coefficient[32], *T*, which measures the degree of similarity between a query and target structure as follows:

$$T = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

Where *A* represents the query fingerprint and *B* the target fingerprint, N_A and N_B represent the number of "1-bit" in *A* and *B* respectively and N_{AB} represents the "1-bit" shared in both. To ensure consistency, analogs were transformed to the same canonical form as the query and filtered if the *T* score was below 80% threshold (score ranging from 0-1 transformed to percentage). This conversion allowed us to further reduce discrepancies between SMILES conversion while identifying chemical analogs. We reasoned that by selecting close analogs with an 80% similarity, as opposed to the suggested 85% threshold by Patterson et al. study[33], we may be able to retain useful moieties related to antimalarial activity (such as functional groups), while maintaining a high degree of chemical similarity between the query and analog compound.

Although a target may have been reported for a given compound, the query compound might be similar but not identical and our goal was to capture these associations. To identify scaffold families, we constructed a similarity network (including the query dataset and searched analogs) by assigning nodes to chemical entities and edges between entities with a Tanimoto coefficient above 80%. With this network we were able to identify query drug-targets from analogs with known mechanisms of action or gene target, elucidating relationships that could be obscured otherwise.

1.4 Drug target identification

In order to provide drug-targets clues, we acquired and incorporated several large datasets with annotated compound-target pairs and created a module to search these sets for compounds closely related to query molecules (**Supplemental Figure 1.1C**). We used the Novartis Chemogenetic Library[34] assembly consisting of 4,185 compounds annotated with their primary mammalian gene target, an in-house, consolidated file of 163 validated antimalarials originating from phenotypic screens with their drug target, a collection of 218 licensed well-validated therapeutic drugs extracted from well-established databases such as IUPHAR/BPS pharmacology database[35] or the FDA approved drugs list[36], and a set of 157 known antibacterial inhibitors[37] from which seven (azithromycin, mupirocin, dapsone, sulfalene, sulfadiazine, triclosan, sulfamethoxazole) overlap with the antimalarial set. The assembled list can be accessed at https://github.com/winzeler-lab/CACTI metadata folder.

To reduce biases when comparing large datasets with molecules of wide mass range, we implemented a method to partition the query and target dataset by molecular weight (default cutoff of 500 Da) and convert each into fingerprints for subsequent similarity calculations. We then implemented a function to allow selection of fingerprint conversion algorithm (RDKFingerprint or GetMorganFingerprintAsBitVect) and selection of binary vector size, to prevent biased structure pattern identifications. Once fingerprints were calculated, we created a NxM matrix, where N and M is the number of fingerprints in the dataset and calculated Tanimoto coefficients for every pair, followed by clustering molecules based on their score (default 0.8).

1.5 CACTI, a large-scale chemical compound analysis tool

As a final step we aimed to integrate the chemical querying and chemical comparison approaches to construct a compound analysis pipeline that will systematically 1) identify index identifiers for a set of given compounds, 2) consolidate relevant information known for index identifiers and identify close chemical analogs, and 3) identify scaffold families to provide target hypothesis with collected index annotations, analogs and known drug-target information. (**Figure 1.1**). The pipeline, CACTI, is accessible at https://github.com/winzeler-lab/CACTI. Shortly, we built the customizable pipeline with python version 3.12 currently executable with UNIX, allowing for single or bulk queries and the selection of analyses and parameters to be performed. After submission validation of the query SMILES set (**Figure 1.1A**), the first step in the pipeline was to use the chemical and biological space exploration method to obtain the multiple identifiers from which a single compound is known, including IUPAC and common name(s), as well as peer-reviewed publications, deposited datasets, and patents associated to each common name (**Figure 1.1B**). To identify analogs in the public domain that might have a known target, we then retrieved the list of similar scaffolds to the ones of interest (**Figure 1.1C**) with at least 80% similarity. This step allowed us to increase the likelihood of accurately uncovering drug-target pairs or providing prediction starting points.

To better understand the query molecules and relationships within the dataset and public domain using the acquired genomic and bioactivity annotations, the next step in the pipeline includes the chemical comparison methods. SMILES for the query set and analogs are transformed to binary fingerprints (**Figure 1.1D**) to allow comparisons between them, and clustered together to find core scaffolds that are similar to at least 80% to each other (**Figure 1.1E**). Although this step seems redundant after querying for similar structures in **Figure 1.1C**, performing this extra

validation was valuable to discover if core scaffolds from the analog subset is shared by two or more query compounds. Lastly, to predict a drug-target under the molecular similarity principal (**Figure 1.1F**), scholar references, known mechanisms of action or gene targets, and bioactivity data associated to compound members within each cluster are compared. It is important to mention that in case of compound analog reacting noncovalently against a biological target yet considered active, typically referred as pan-assay interference compounds (PAINS), target hypotheses may be misguided for a query compound based solely on the bioactivity annotations. Thus, careful inspection may be needed when evaluating members from the cluster and additional annotations may need to be considered.

1.6 Results

To facilitate the exploration of the chemical and biological space for large datasets, and to provide a better understanding of chemical structures and their impact on a biological system, we constructed a drug-target prediction tool using data mining and chemoinformatic techniques. Our overall objective was to automate the tedious molecule-by-molecule searching that occurs when hits from a high throughput phenotypic screen are initially evaluated, and to perform this search in a comprehensive, unbiased fashion, looking for information not only on the query molecules but also on closely-related analogs. Because we noted that molecules might have different common names in the literature, we also sought to systematically uncover synonyms. For example, the new imidazolopiperazine antimalarial named ganaplacide that is in phase III clinical trials has been known as KAF156, and GNF156 and substantial work has been published on the closely related scaffold named GNF179 which differs from KAF156 by a single halogen substitution. Our motivation was thus to provide a tool that would allow the user to focus energy

on phenotypic screening hits that might have novel mechanisms of action and thus avoid compounds with a well-established mechanism or known drug resistance liabilities. In addition, we wished to create a tool that could alert the user to relevant information from work on other species. If a compound is a dihydrofolate reductase inhibitor in humans, it is likely also a dihydrofolate reductase inhibitor in malaria parasites[38].

As described in the **methods**, this tool, which we have named CACTI consists of 3 modules, was coded into python and accepts queries in the forms of tabular or comma-delimited files containing the SMILES of query molecule(s), as well as additional field columns such as the standard name, if desired. The output consists of Excel files from each selected analysis (module) with a comprehensive report for all query molecules with synonyms and scholar evidence mined from the selected knowledge databases including ChEMBL, PubChem, BindingDB and PubMed. The output also includes a complete list of close analogs to query compounds and their similarity scores from the network construction module. Lastly, it includes a report from a clustering module in which query compounds are matched to similar molecules from the annotated target set. These last two reports can be exported to external tools for network visualizations, such as Cytoscape[39]. CACTI can be accessed from the public GitHub repository and executed by creating a local copy and following instructions indicated in the repository.

1.7 Case study: assessment of drug-target querying using the Pathogen Box dataset

In order to assess the performance of the tool, we applied it to the Pathogen Box dataset[40] from the Medicines for Malaria Venture (**Figure 1.2A**). This set is a collection of 400 diverse drug-like molecules, with an average molecular weight of 374.13 Da, active against several

neglected tropical diseases. This set was physically assembled by the Medicines for Malaria Venture and shared with users working in the neglected disease space to provide drug-like starting points for oral drug discovery. Accordingly, only 32 compounds failed to comply the standard Lipinski Rule of Five[41]. The set includes 26 positive controls from various scaffold families, 24 of which are part of known drug-target pairs. Controls include well known drugs like the antimalarials doxycycline, primaquine and mefloquine[42], the antibacterial rifampicin[43], the antimicrobials pentamidine, nifurtimox and fluoxetine, the anti-schistosomiasis medicines, praziquantel and mebendazole. For these controls, we often found too much information, and these were thus excluded from further consideration. The remaining 374 compounds (**Supplemental Table 1.1**) are ones that have reported activity in published phenotypic screens for neglected tropical disease, with the majority being antimalarial or antimycobacterial compounds[44-46] (**Figure 1.2B**). Chemical properties for the 400 compounds and biological activity in multiple parasites and pharmacokinetic measurements, including cellular toxicity and pharmacokinetic measurements in HepG2 cells, can be found in ChEMBL-NTD repository (https://chembl.gitbook.io/chembl-ntd) set 21.

1.7.1 Module 1 identifies 4,315 synonyms for the Pathogen Box compounds

A good literature search can prevent months of wasted effort, but searching biological databases such as PubMed is hampered by the fact that compounds frequently change names in biological publications. We assessed if we could find synonyms for the compounds in an automated way. Searching using CACTI identified 4,315 synonyms for compounds in the Pathogen Box, that were not included in the initial Pathogen Box description, with an average of 11 names per compound (**Supplemental Table 1.1**, and **Figure 1.2C-D**). For example, we found

10 synonyms for compound MMV006741, including GNF-Pf-4484, Maybridge3_001644, and BRD-K92165166-001-01-3, among others. Likewise, the automated synonym search revealed that one trivial name for MMV676600 is danusertib, a known pan-aurora kinase inhibitor[47]. Overall, we found 69 different names for MMV676600. These data also provided clues about the provenance of the compounds. The search revealed substantial overlap between the GSK TCAMS (Tres Cantos Anti-Malarial Set) library[48] (ChEMBL-NTD (https://chembl.gitbook.io/chembl-ntd), set 1) with 164 compounds present in both sets, 74 compounds overlapping with the St. Jude Biomedical library[49] (https://chembl.gitbook.io/chembl-ntd, set 3) and eight that were members of the Novartis GNF library[50] (https://chembl.gitbook.io/chembl-ntd, set 2). The automated search also revealed members from the Pathogen Box that are readily accessible to the scientific community. We found 206 compounds with suppliers' identification numbers for various commercial sources including AKOS, Zinc and MCULE. In general, module 1 was able to capture and identify diverse names from published studies and public libraries, as well as provide the unique identifiers for selected databases including PubChem (CID identifier) and ChEMBL (CHEMBL identifier).

1.7.2 Module 1 identifies new citations, datasets and patents for Pathogen Box molecules

No literature citations were provided with the initial Pathogen Box dataset. Therefore, we next used CACTI to comprehensively identify patent and literature citations using not just the MMV names but all synonyms (**Figure 1.2E**). Unsurprisingly, searching with the MMV names often produced literature that described the Pathogen Box. However, searching with synonyms provided many more citations on the compounds. For example, MMV675997, is associated with 4 pieces of evidence that were deposited separately under several chemogenomic databases with different

synonyms (MMV675997, CHEMBL1094051, BDBM50313769) (**Supplemental Table 1.1**). Searching with these synonyms showed that this compound, which has a peptoid backbone and a nitrile P1′ warhead came from a library predicted to contain *T. brucei* Metacaspase enzyme inhibitor[51]. Likewise, the kinase inhibitor, danusertib (MMV676600) is associated with a total of 1,195 pieces of evidence with citations indicating a possible role in inhibition of *Toxoplasma gondii* calcium-dependent protein kinase 1 CDPK1 (calcium-dependent protein kinase 1), as well as 100 pieces of patented evidence. From these, a CACTI PubMed search using PHA-739358, the additional synonym of MMV676600, yields three more literature references compared to MMV676600 (**Supplemental Table 1.1**).

The literature search also showed that some of the compounds have activity in multiple species. For example, Murphy and colleagues[52] discovered that MMV676050 and MMV676182, both tested against cryptosporidiosis in the Pathogen Box activity profiling, are potent inhibitors targeting CDPK1 *T. gondii* with known crystal structure (PDB ID 3N51[52]). Similarly, the literature search revealed that 162 members of the Pathogen Box showed evidence of a target validation in other models, such as cancer cell lines. For example, the chemotherapeutic milciclib (PHA-848125), currently in clinical trials, is a known dual cyclin-dependent and tropomyosin receptor kinase inhibitor[53, 54]. Overall, we found 30,425 new pieces of literature, 3,356 patents and 2,182 datasets associated with the Pathogen Box compounds. Though exploring the literature for different names may be straightforward for a small-scale study, as the study size increases, it becomes more difficult to capture these nuances. Therefore, the implemented method of acquiring and cross-referencing all known identifiers to retrieve data from the several chemogenomic databases provides a more detailed and complete report.

1.7.3 Module 2 identifies 12,383 new closely related compounds in the public domain that can be used for SAR by inventory

An important step for compound evaluation and testing is determining if there are closely related analogs available in the public domain that can be sourced for testing. Performing a Tanimoto similarity search for the 374 uncharacterized Pathogen Box members resulted in 12,383 close analogs (T = 90%) from the selected chemical knowledge databases (**Supplemental Table 1.2**, and **Figure 1.2D**). In contrast, only eleven compounds (MMV393995, MMV676269, MMV676442, MMV021057, MMV202553, MMV688754, MMV099637, MMV676050, MMV676064, MMV676204 and MMV676398) out of the 374 had no obvious analogs in the public domain under this criterion. Most analogs identified belonged to PubChem (12,246 analogs), BindingDB (112), or ChEMBL (25) databases. On average, 34 closely related analogs were identified for each member of the query set with an average similarity of 94.44%. For example, we found 55 analogs of MMV019721, a recently discovered acetyl-coenzyme A synthetase *Plasmodium* inhibitor[55]. Many of these analogs are commercially available.

Interestingly, we observed eight closely-related analogs that were identified twice when querying for the 374 Pathogen Box compounds (**Supplemental Table 1.2**, and **Supplemental Figure 1.2**). Finding a close analog for two different query compounds suggest the presence of similar core scaffolds and likely similar biological targets. For example, the compounds MMV595321 and MMV676477, share five different analogs (CID's 136636721, 136636828, 156278160, 156278182 and 156285860), most of which are part of an antiparasitic inhibitor optimization effort (patent WO-2021077102-A1). Another analog (CID 44526919), related to

MMV023969 (TCMDC-134161) and MMV024035 (TCMDC-134227), inhibits *P. falciparum* growth by up to 97% (IC_{50} < 2 μM)[56]. Thus, the related Pathogen Box compounds could be attractive drug starting points.

1.7.4 Module 3 for new target discovery hypotheses —guilt by association approaches

One of the most resource-consuming steps in drug discovery is target deconvolution. To create target predictions, we utilized the information collected from previous modules and clustered the 374 Pathogen Box compounds and 156 close analogs, with sets of 4,716 molecules with known mechanisms of action to create hypotheses about the function of Pathogen Box compounds. The "known target" set was derived from multiple sources, including 4,185 compound-target pairs assembled by Canham et al.[34], 150 known antibacterial drug-target pairs[37], and a collection set of 381 antimalarial compound-target pairs extracted from established databases such as IUPHAR[35] and PDB Ligands[57], in addition to validated target-ligand pairs originating from the TCAMS and GNF Novartis Malaria Box phenotypic screening libraries. Although the assembled library of known target pairs was enriched for antimalarial inhibitors, many have activity across the parasitic and bacterial diseases included in the study, such as atovaquone and doxycycline. It is worth noting that targets for 152 Pathogen Box compounds (**Supplemental Table 1.1**) were already known and were already included in the known target set. Clustering the 374 and their 156 close analogs with 4,716 known targets resulted in 77 clusters using a similarity threshold of $T = 80\%$ **(Supplemental Table 1.3)**, with an average of 4 compounds per cluster, and 264 singletons (**Figure 1.3A**). Out of the 77 clusters, twenty-five had two or more Pathogen Box molecules and 71 clusters contained one or more MOA/target or putative target annotations. From the 71, 20 have at least one Pathogen

Box compound with unknown or unclear mechanism. We did not find a MOA/target prediction for six clusters (**Figure 1.3B**).

To assign targets for the Pathogen Box compounds, we inspected members from each cluster and assigned a potential target/function for the "unknown" Pathogen Box compound based on their counterpart target evidence. From this, we found targets appear repeatedly at rates higher than expected (hypergeometric mean function = $4.4x10^{-3}$ – $2.6x10^{-48}$) (**Figure 1.3C**). For example, three clusters (55, 56, 212) contain predicted dihydroorotate dehydrogenase (DHODH) inhibitors and six contain predicted cytochrome b inhibitors, of which cluster 62 have one Pathogen Box compound (MMV687807) with no known reported target in *Plasmodium* or *Mycobacterium*. We also found 9 clusters whose members contain an analog or known target set compound from two or more different annotations. This includes cluster 338 where close analogs to MMV687700 are known salicyl-AMP inhibitor and other members from the known drug target set are known human adenosine receptor agonists. Despite this apparent incongruity, all compounds in cluster 338 contain the basic adenosine scaffold and are adenosine analogs. Another example includes cluster 177 where MMV688283 and MMV687246 (a *Pf*CDPK5 inhibitor[58]) are clustered with TCMDC-138293, a known *Plasmodium* DHODH inhibitor. This may suggest some polypharmacology or potential weaknesses in the automated annotations.

We also found cases where the predicted target was assigned based on evidence from a different parasite genus to the one initially tested. The diazine scaffold family (cluster 30) has two compounds that have demonstrated activity against parasite and mammalian methionine aminopeptidase-1b, an enzyme predicted to catalyze the removal of the N-terminal initiator

methionine during protein synthesis in parasites and mammals (MMV084997/GNF-Pf-359). The group also contains the uncharacterized, anti-kinetoplastid Pathogen Box compounds MMV658993 and MMV658988. This association creates a hypothesis that MMV658993 and MMV658988 may target methionine aminopeptidase-1b in kinetoplasts.

1.8 Challenges and limitations

Neglected diseases (parasitic diseases, and to some extent, rare diseases), attract less commercial interest and significantly less overall funding than other diseases, such as cancer or diabetes. On the other hand, much of the small molecule data for well-funded diseases such as cancer lies in the private, well-curated databases that are maintained by pharmaceutical companies. These databases are often one of a company's biggest intellectual property assets. In contrast, for neglected disease, a larger proportion of the drug discovery data will be generated by academic researchers or public-private partnerships where much of it will ultimately be placed in the public domain. The availability of well curated public screening datasets (e.g. such as the Pathogen Box dataset) creates both opportunities to make new discoveries (e.g. for drug repurposing and target discovery) as well as challenges that relate to the dispersed and disorganized nature of the data. To make the task of assessing diverse data more accessible to neglected disease drug discovery researchers, we constructed a tool to pragmatically assess the major chemical databases and identify all data available for a query compound, as well as similarities between small molecule inhibitors in any system. With this approach a comprehensive report is generated and may be used to redirect laboratory resources in a more focused way, in addition to reducing the amount of work needed for drug-target or chemical optimization efforts.

Our approach has its limitations. First, our algorithm makes no assessment of literature quality. A manuscript reporting that a given compound binds a target may not rigorously assess the quality or strength of the binding. This may happen when a researcher develops a biochemical assay for a target and then tests a library, such as the Pathogen Box library, against this target: In this case the best compounds from this exercise may be reported as potential inhibitors of the target in a publication. Our computational algorithms cannot catch these nuances and are out of scope; although a thorough report on a particular query set will be obtained, a human review is needed to revise and confirm the veracity/strength of acquired evidence. Perhaps not surprisingly, the data-querying strategy across multiple chemogenomic repositories is especially helpful when querying for small molecules that have been evaluated for activity against multiple disease indications. For example, compounds advancing to clinical-trials or those having a validated drug-target, are query entries that when searched provide the most complete information.

Another limitation is that our approach somewhat relies on the assumption that a compound will have the same target in *Trypanosoma cruzi* as in *Toxoplasma gondii*, despite the two species belonging to different eukaryotic phyla. Nevertheless, previous studies have highlighted that drug-target pairs tend to be conserved across species. A compound like methotrexate likely acts against dihydrofolate reductase in human and in malaria parasites. Cladosporin targets lysyl-tRNA synthetase in both yeast and malaria parasites[59]. There are hundreds of other examples of target conservation across species. Species selectivity thus comes from natural or engineered specificity, as well as innate ability of different species to detoxify compounds. On the other hand, just because a compound targets the electron transport chain in malaria parasites doesn't mean that it targets the electron transport chain in *M. tuberculosis* (e.g. such as in ELQ-300, with unreported target in tuberculosis) and predictions and caution is needed. As a future improvement to the pipeline,

similar to gene ontology strategies (GO terms), querying and comparison of medical subject headings (MeSH terms) associated to compounds of interest (and analogs) can be implemented to mitigate the need to rely on similar structure function assumptions. This could increase the confidence of CACTI target identification hypotheses.

Another concern is that we relied on chemical scaffold clustering: This chemical comparison depends on the chemical fragment partition that is initially selected, resulting in potential arbitrary cluster assignment. In addition, clustering is less suitable for natural products, which tend to have more complex structures than smaller molecules. Increased size means partition methods favor fragment partitions that break down molecules and ignore the R groups that may be vital for the pharmacological effect of the natural product. Thus, although the provided scholarly report will be helpful, the clustering approach as scripted is less suitable for their evaluation. Nevertheless, we found that close analogs with similar drug-target matches did tend to group together even if the larger group often split into more than one cluster. We are confident that the established prediction method provides an initial hint for further target validation efforts.

A limitation with this approach includes the predictive component of the analysis. Though it is accepted that similar molecules tend to have similar mechanisms of action or drug-targets, experimental validation is needed to confirm binding to the desired target. The assumption that one small molecule will have just one target is more fraught as compounds get larger and where two different pharmacophores (e.g. an ATP-like molecule and a naphthalene) may be incorporated into one small molecule. In addition, for promiscuous inhibitors, like kinase inhibitors, our approach may provide limited resolution.

Another concern is that one component of this tool relies completely on network connectivity and database server availability (e.g. open query-request for selected databases) to query and retrieve literature and analogs from the public domain. To partially address this need, a future avenue includes the use of a periodic "database dump" that can be locally accessed and queried in lieu of internet access, when connectivity is unreliable or inaccessible.

Overall, we believe the identification of literature evidence as well as close analogs obtained through CACTI will aid researchers in early stages of drug discovery pipeline, and consequently allow more work on structure-activity relationship analysis. Similarly, the target prediction module implemented in this tool will serve as starting points for subsequent drug-target validation efforts and support the translation of genomic data into effective new drugs through the comparison of scaffold families. This automated tool shows a promising approach to quickly investigate multiple chemical entities in a single query, and prioritize hits for further exploration, especially in academic settings where compound management and resupply is limited.

1.9 Acknowledgements

Chapter 1 is a partial reprint of the material as it appears in Journal of Cheminformatics, 2024. Karla P. Godinez-Macias and Elizabeth A. Winzeler, "CACTI: an in silico chemical analysis tool through the integration of chemogenomic data and clustering analysis". The dissertation author was the primary author of this paper.

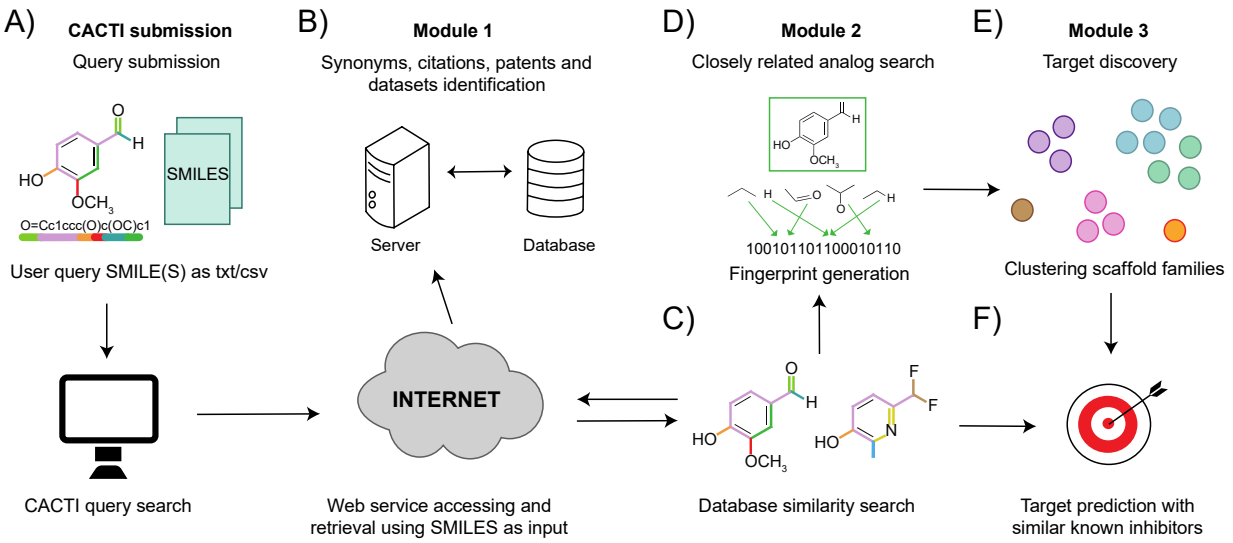


Figure 1.1 CACTI pipeline workflow

Pipeline workflow steps are illustrated by modules where **(A)** represents the user SMILES set as input in a tabular or comma delimited, **(B)** refers to the database querying and retrieving steps using SMILES to identify synonyms, citation evidence, patents, membership in datasets and additional information for target prediction. Module 2 refers to **(C)** the identification of close analogs in chemogenomic databases to the query set, and **(D)** transformation from SMILES to binary fingerprint for scaffold comparison. In module 3, **(E)** clustering chemical entities based on similarity measures is performed and **(F)** target predictions are completed combining information from previous modules and validated known-target datasets.

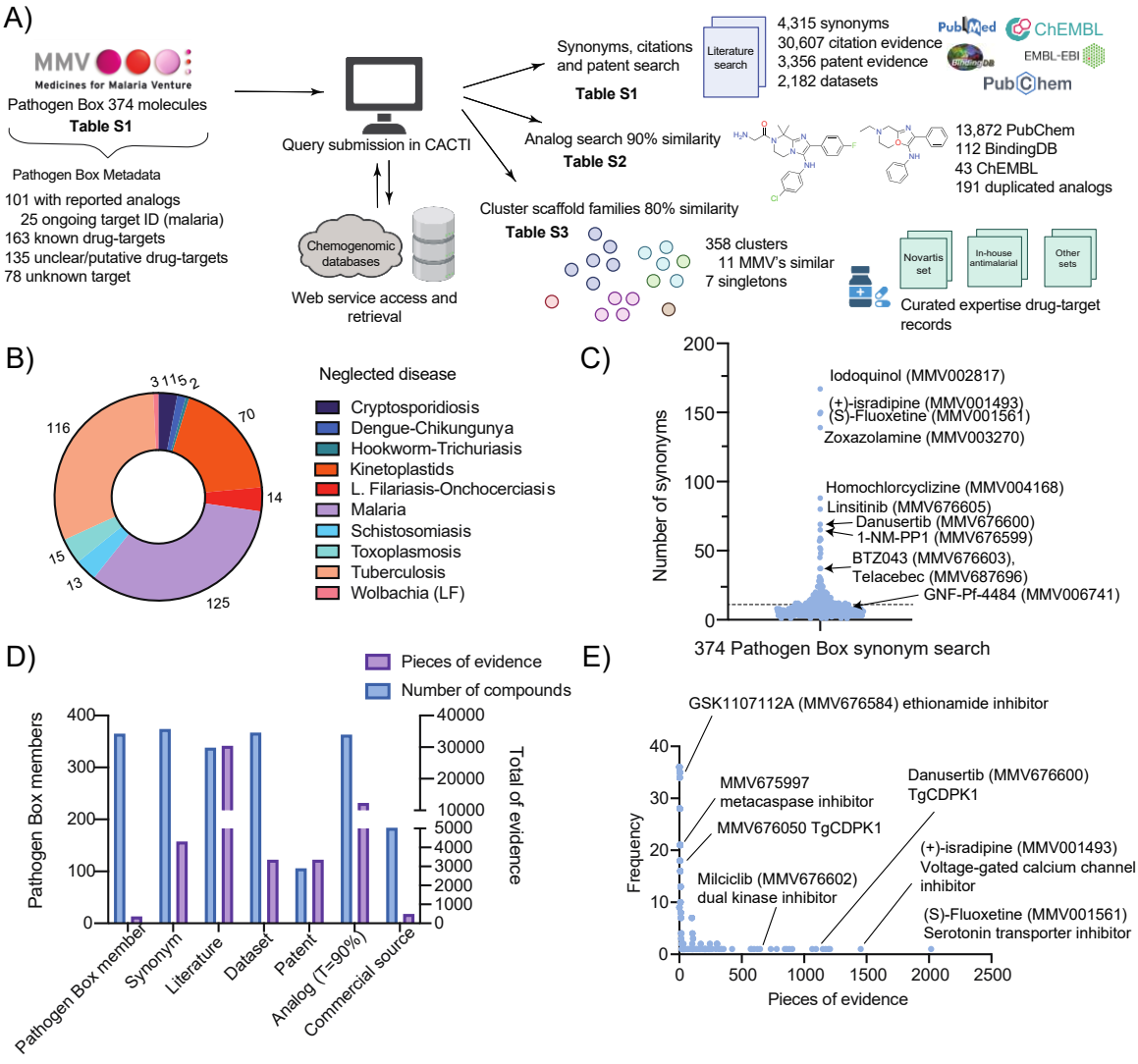


Figure 1.2 Identification of synonyms, references, and close analogs for 374 Pathogen Box compounds

(A) Pathogen Box querying workflow. Results obtained from querying 374 Pathogen Box compounds against the pipeline. Supplemental tables related to each analysis are indicated along with the results from each module. **(B)** Membership of selected compounds. Number of compounds active against each neglected disease according to the Pathogen Box reference biological data[40]. **(C)** Synonyms identified. Number of names given for each Pathogen Box compound across different studies. Well established drugs belonging to the Pathogen Box are shown along with the common drug name. A dotted line denotes the average (n = 11) of synonym names per compound. **(D)** Statistics for identified information per compound. Identified categories for the 374 compounds are shown in the X axis. Left Y axis shows the number of Pathogen Box compounds found per category and the right Y axis shows the number of pieces of evidence per category. **(E)** Pieces of literature found per compound. Scatter plot showing the total pieces of evidence for each compound (X axis) against the frequency (Y axis). Known inhibitors with their inhibitory protein/enzyme are shown.

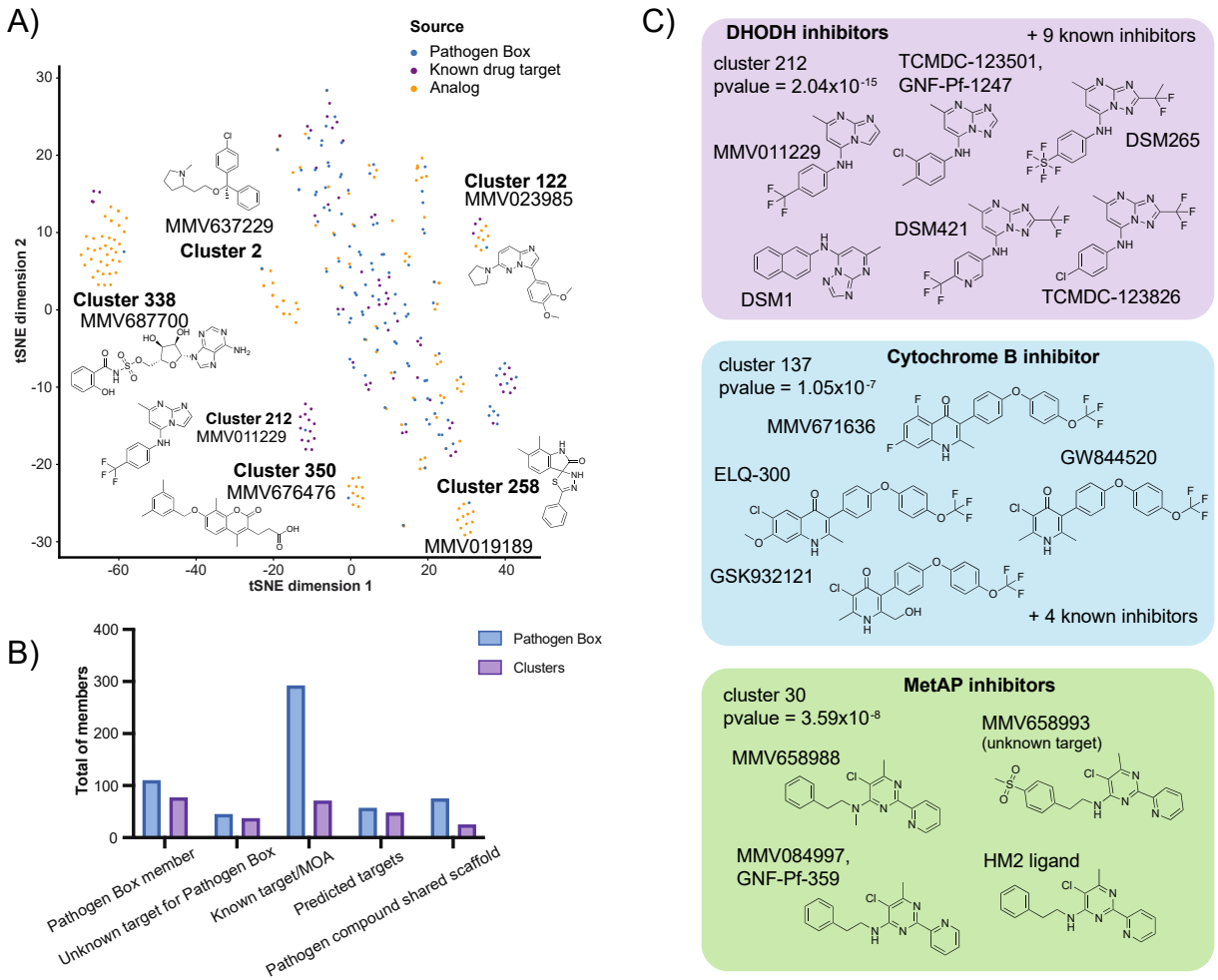


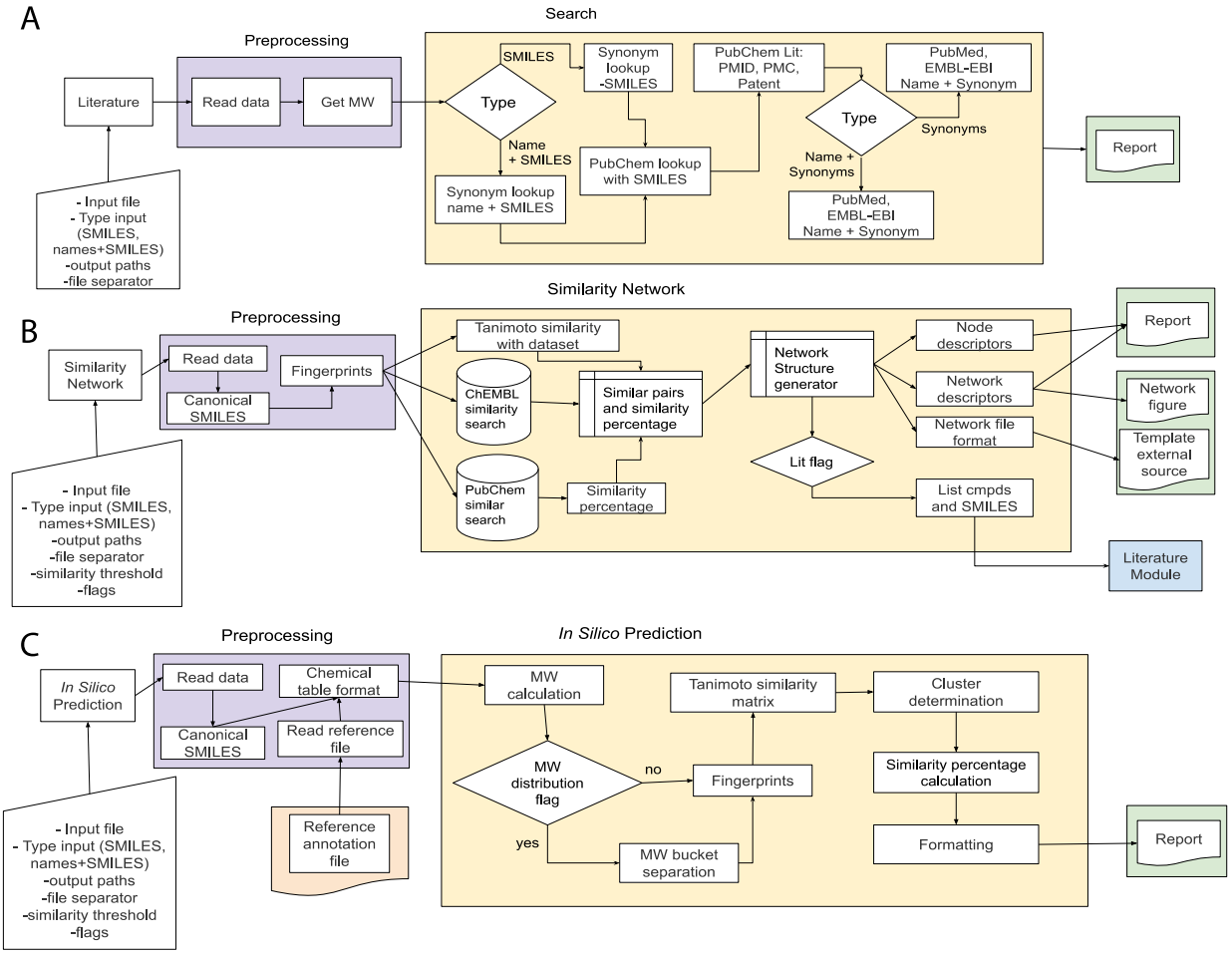
Figure 1.3 Chemical clustering and target prediction

(A) Scaffold clustering. Structural similarity calculations with Pathogen Box library and known drug targets, as determined with RDKIT[31] fingerprints. Clusters were assigned using 80% Tanimoto similarity threshold. Visualization of chemical space was performed using scikit-learn t-SNE function. **(B)** Scaffold clustering statistics. Total number of Pathogen Box members (light blue) and clusters (light purple) identified for the different categories shown in *X* axis. Total of known drug-target annotations, and predicted drug-targets, among all clusters are included in the summary categories. **(C)** Overrepresented clusters with known target. Examples of clusters with drug-targets found at rates greater than expected according to the hypergeometric mean function. Structures for cluster members are included.

Table 1.1 Number of records per database.

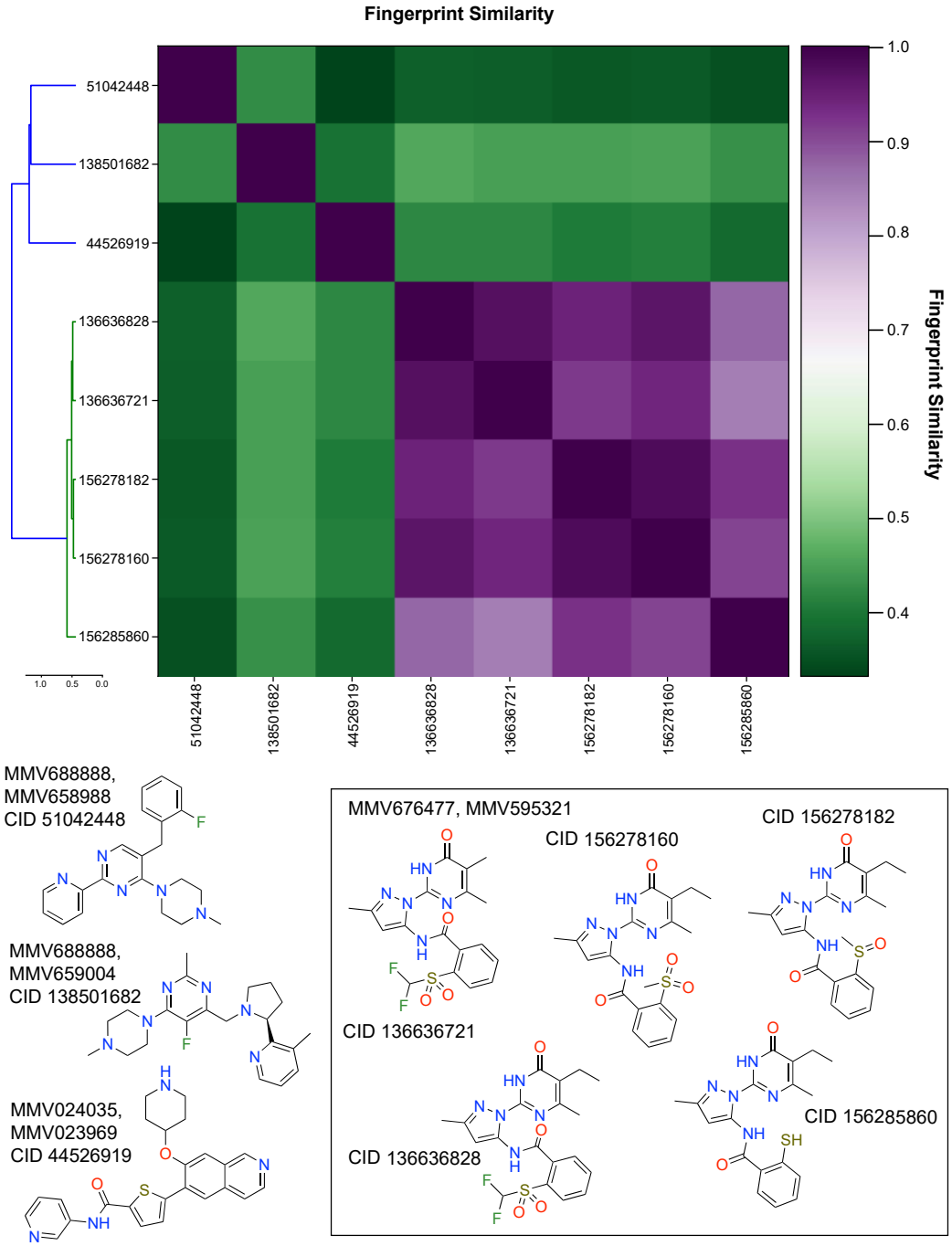
Approximate number of records by entity type for selected chemogenomic and literature databases.

Database	Chemical Data	Experimental Data	Literature	Patent	Compound-Pathway/Gene
ChEMBL	22,399,743	20,334,684	88,630		15,398
SureChEMBL				> 17 M	
PubChem	115,036,406	304,160,942	38,884,316	43,059,414	Pathway: 240,631
BindingDB	1,164,246	2,707,699	39,818		9,023
EMBL-EBI			34,135,964		
PubMed			~ 34 M		



Supplemental Figure 1.1 Modules for drug target prediction

Module workflow for **(A)** literature search, **(B)** similarity network and analog search, and **(C)** prediction. Trapezoidal boxes show input data types of choice for each module, purple boxes represent preprocessing steps to each section. Yellow boxes delimitate the steps to complete a module's task, where diamonds are decision steps and rectangles a function to be performed. Green boxes are figure or results report generated.



Supplemental Figure 1.2 PubChem duplicated analogs fingerprint similarity

Scaffold similarity comparison for the 8 PubChem (CID) close analogs for two different Pathogen Box query compounds. Fingerprint similarity scores were calculated using RDKIT Tanimoto similarity function. Left panel shows the hierarchical clustering for each CID based on the calculated similarity, central panel shows the similarity intensity color coded according to the left panel gradient, where low similarity (0-0.5) is represented by green and high similarity (0.5-1) is represented by purple. Structures for analogs are shown, including the source Pathogen compound.

Chapter 2

Systematic in vitro evolution in *Plasmodium falciparum* reveals key determinants of drug resistance

2.1 Introduction

As drug resistance remains a major concern in controlling malaria, an infectious disease caused by protozoan *Plasmodium* parasites, detection of resistance is of vital importance. Because large-scale phenotyping of parasites for drug resistance is impractical and results of therapeutic efficacy studies are limited, a major goal of sequencing clinical isolates is to identify the emergence of genetic markers of drug resistance, thus providing an early warning to inform region-specific malaria treatment policies. As a result, surveillance efforts have now placed whole-genome sequences of more than 20,000 isolates of *Plasmodium falciparum*, the most common and virulent malaria parasite, into the public domain[60]. A key challenge in these efforts is distinguishing functional variants, which drive the observed phenotype, from passenger mutations, i.e. those that do not confer phenotypic change. While previous studies have identified some clinical resistance markers, the identification of novel resistance alleles either in laboratory-adapted or field isolates remains inefficient.

Since the earlier use of genetic crosses to determine key genes mediating parasite resistance to pyrimethamine, chloroquine, and other antimalarials, two alternative approaches for understanding drug resistance have emerged: population-based association studies of sequenced isolates, and in vitro evolution of resistant parasites[61]. An advantage of in vitro evolution studies is that because relatively few mutations emerge during drug selection, variants with a high probability of conferring resistance can be discovered through careful genome analysis of only a handful of independently derived drug-resistant clones. As a result, hundreds of laboratory-evolved organism genome sequences have been published. While studies of malaria parasites are among the most advanced for this type of work, the approach is also becoming widespread for other important microbes, including *Candida albicans*[62], *Mycobacterium tuberculosis*[63], and the model organism *Saccharomyces cerevisiae*[64].

To explore the potential of in silico approaches for predicting antimalarial resistance mediators by leveraging insights from in vitro evolution, we comprehensively analyzed whole-genome sequences of 724 *P. falciparum* compound-selected clones, including 210 created specifically for this work. We performed meta-analyses across compound-selected mutations to identify characteristics of functional variants associated with resistance. In addition to revealing new resistance alleles and genes, our analysis identified protein and genome features associated with drug resistance and specific protein domains that consistently yield resistance. We found that mutations in non-coding regions or the non-core genome (subtelomeric and internal regions comprised of hypervariable gene families such as *var*, *rifin*, and *stevor*[65]) seldom contribute to resistance phenotypes. Our dataset provides a starting collection for algorithms that can identify genomic changes that are likely associated with drug resistance and presents insights for distinguishing functional from nonfunctional variants in forward genetic approaches.

2.2 Results

2.2.1 Compound-selected clones contain few mutations

We first sought to assess whether most of the nonsynonymous mutations in compound-selected clones were functional. Using laboratory resources of the MalDA Consortium[66], we compiled a set of 724 whole-genome sequences from in vitro evolved *P. falciparum* strains and their isogenic parents (**Supplemental Table 2.1**). Blood-stage parasites were exposed to one of 118 small molecule growth inhibitors, ranging from tool compounds identified by phenotypic screening as reviewed in [67], licensed antimalarials, and compounds in the developmental pipeline (**Table 2.1**), for periods ranging from weeks to more than two years. These 118 compounds were collapsed into 95 groups based on shared chemical groups. Although some datasets were

previously published, this work contains the meta-analysis of our entire repository of previously reported samples, samples downloaded from the NCBI Sequence Read Archive (SRA), and 210 new genomes of parasites resistant to 33 compounds (**Supplemental Table 2.1**).

Clones with phenotypic data ($n = 448$) were on average 56-fold (ranging from 1.1- to 2,654-fold) more resistant to their respective compound than the matched parent (**Supplemental Table 2.1**). Although sequenced in multiple locations, genome sequences were acquired using similar paired-end short-read whole-genome sequencing (WGS) methods on Illumina platforms, with an average coverage of 135×. Most clones were selected from the Dd2 ($n = 340$), Dd2-polδ ($n = 35$), or 3D7 ($n = 328$) strains, with a few derived from 7G8, NF54, or W2 (see **methods**). In some cases, multiple clones were isolated from the same selection flask and proved genetically identical; these were treated as technical replicates.

To standardize identification of mutations that evolved during selection, we aligned sequencing reads to the *P. falciparum* 3D7 reference genome. To identify high-quality variants, we applied hard filtering on single nucleotide variant (SNVs) and insertion-deletion variants (indels) according to GATK recommendations[68]. Because the *P. falciparum* genome is haploid in asexual blood-stages and most samples were expected to be clonal, we retained variants for which 90% or more of read calls mapped to alternate (non-reference) alleles. Variants were then assigned predicted effects using the annotation tool SnpEff[68] (**Supplemental Figure 2.1**). Because each evolved sequence came with an isogenic parent, we subtracted variants present in the parent that had not emerged over the course of compound selection. Altogether, 5,560 SNVs and indels were identified across the dataset (**Supplemental Figure 2.1**), with an average of 17 mutated genes shared within individual compound chemotype groups. Of these variants, 4,105 were in the "core" genome, which excludes genomic regions with repetitive sequences[65] (**Supplemental Table 2.2**).

Each clone contained, on average, 7.65 mutations relative to its parent, of which an average of 1.2 were missense mutations in core regions. Given that there are 5,247 core genes in the genome, the data show remarkable specificity and suggest that a large proportion of the identified mutations play some role in compound resistance.

2.2.2 Compound-selected mutations differ from passenger mutations

The characteristics of our compound-selected set of SNV/indel mutations were compared to a set of control SNVs/indels that differ between two non–compound-selected clonal strains. Removing technical duplicates from the compound-selected set reduced the number to 2,628 independently derived mutations, of which 1,448 were in the core genome (**Supplemental Figure 2.1**). For the control set, we compared two laboratory strains, Dd2 and 3D7, originally derived from Southeast Asia and sub-Saharan Africa, respectively. Comparisons between Dd2 and 3D7 sequences using an identical pipeline and filters resulted in 48,364 variants, of which 38,064 were core variants. Overall, our compound-selected variants were proportionally more likely to be missense or frameshift mutations (**Figure 2.1A**), while variants distinguishing the control clonal isolates were more likely to be synonymous. In terms of nucleotide changes, compound-selected mutations were more likely to be G to T transitions (**Figure 2.1B**). Our data suggest that the compounds used in selections were generally not mutagenic, although there could be exceptions. The distribution of amino acid changes was also distinct from that of the control set; transitions were more likely to be from a small to a bulky amino acid such as phenylalanine or tyrosine (**Figure 2.1C**). Finally, we examined whether core mutations were likely to be in a recognized protein domain. Around half of the core missense mutations (n = 685) in the compound-selected dataset were located within an InterPro protein domain. By contrast, less than 10% of the 5,969

core missense control mutations (n = 555) were in a recognized protein domain (**Figure 2.1D**). These differences in distributions were significant ($P < 0.001$, hypergeometric test), highlighting that a substantial proportion of compound-selected mutations are likely to play a functional role and are unlikely to be the result of random genetic drift.

2.2.3 Copy number variants frequently drive compound resistance

Copy number variants (CNVs) have been shown to mediate clinically relevant drug resistance phenotypes in malaria parasites[69]. To identify CNVs in our WGS dataset, we calculated differences in read coverage for core genes and compared coverage to that of a panel of untreated controls assembled from parent clones, depending on whether reads were derived from the 3D7 or Dd2 strain background (**Supplemental Table 2.3**, and **Supplemental Figure 2.2**). To identify the amplified or deleted sequence region comprising the CNV, we identified groups of four or more genes with $\log_2 > 0.4$-fold change in coverage relative to their parent strain. This resulted in 1,168 potential amplification events. Using a Kruskal-Wallis test comparing copy ratios of genes within the amplification boundaries to those averaged across the corresponding panel of controls, we found that 271 of the 724 clones harbored at least one of 420 CNVs with $P < 0.0001$. These CNVs included amplifications of known targets, such as *pfpi4k*, *pfproRS*, and the acetyl–CoA transporter, as well as multidrug resistance genes (*pfabci3*, *pfmdr1*) (**Figure 2.2A**). We also identified multiple CNVs, including one on chromosome 14, that had been noted as emerging after long term exposure to artemisinin[70]. Amplifications of GTP cyclohydrolase1 (*pfgch1* on chromosome 12, PF3D7_1224000) were not generally included in this set, as many strains have more than one copy of *pfgch1*[71].

To infer the gene within a CNV that confers a selective advantage, we looked for evidence of a known target or resistance mechanism within the amplified segment or for recurring CNVs and identified the gene that was amplified most frequently across the dataset in overlapping CNVs. *Pfmdr1* was amplified in 47 independent CNV events (**Figure 2.2A**) and harbored 19 independent SNVs, making it a clear driver of resistance. For smaller CNVs, there was more ambiguity. A set of five genes on chromosome 12 was amplified 11 times, but of this set, only *pfprs* (PF3D7_1213800, proline-tRNA ligase) also contained SNVs. In several cases, there was no clear candidate resistance mediator, including a recurring CNV on chromosome 1 and one on chromosome 9. Not all CNVs are related to compound pressure. Many evolved strains contained amplification of an 86 kb region on chromosome 10 that was also found in some parent strains, including 3D7; it may contain unidentified fitness-conferring factors.

Characterization of amplification CNVs by structural variant type and breakpoint properties showed several interesting features. Paired-end read data showed that 179 of 243 independent amplifications were tandem duplications, as opposed to inverted or interchromosomal duplications. We did not find evidence that specific genomic locations are strongly preferred for amplification of driver genes, as many CNVs had different endpoints even for the same compound. For example, CNVs with varying size and breakpoint locations containing *pfmdr1* were found in resistant clones evolved from the 3D7 parent strain, which contains a single copy of *pfmdr1* (**Figure 2.2B**). CNV breakpoints were determined, with nearly single base resolution, using discordant read pairs for 142 high-confidence tandem duplications. These breakpoints showed an enrichment of duplication recombination sites in intergenic regions, which typically have lower sequencing coverage (**Figure 2.2C**). Tandem duplication recombination sites often occurred in highly AT-rich and/or repetitive regions, supporting the hypothesis that long A/T tracks present

throughout the *P. falciparum* genome facilitate microhomology-mediated repair as a major mechanism of increasing copy number[72] (**Figure 2.2D**).

2.2.4 Overrepresented genes affect compound resistance and culture adaptation

Experimental evolution is effective at associating compounds with targets or drug resistance genes if they appear at rates higher than expected by chance. Of the 118 compounds, 47 had a target or resistance mechanism suggested by overrepresented SNVs alone, 14 by both CNVs and SNVs, and 15 by CNVs with a clear driver gene alone (**Supplemental Table 2.4**). For 18 compounds, a resistance gene was determined by additional experimentation. In only 18 cases was no resistance gene identified, indicating a high success rate for experimental evolution (**Figure 2.3A**). On a compound basis, the strongest signatures were for KAE609 (cipargamin), with 12 independent SNVs in *pfatp4* (non–SERCA-type Ca2+-transporting P-ATPase) ($P = 9.98 \times 10^{-41}$; **Supplemental Table 2.4**), and cladosporin, showing three independent CNVs amplifying lysine-tRNA ligase (KRS1) on chromosome 13.

An advantage of this dataset is that it permits the identification of genes that appear repeatedly across disparate compounds, suggesting a multidrug resistance mechanism. Altogether, 128 genes were altered in more than one clone, 53 of which were identified three or more times. The list of statistically significant recurring genes (**Supplemental Table 2.1**) contained multiple antimalarial multidrug resistance genes, such as *pfmdr1*[73] (19 independently derived SNVs/indels and 47 CNVs), chloroquine resistance transporter *pfcrt* (nine independent disruptive variants)[73], and cyclic amine resistance locus *pfcarl* (14 independent SNVs/indels and one CNV)[74]. The list also contained clinically important antimalarial targets, known to acquire resistance mutations affecting inhibitor binding, such as *pfatp4*[75], *pfpi4k*[76], *pfcytb*[77], and *pfdhodh* (totaling 14 CNVs and

seven SNVs)[78]. Another overrepresented gene was *pfap2-g* (PF3D7_1222600), which encodes a transcription factor involved in commitment to gametocytogenesis[79], as well as the Rap guanine nucleotide exchange factor *pfepac*, both known to mutate as a result of long-term culture in the absence of drug pressure[80]. Several large, conserved proteins were overrepresented, including PF3D7_0619300, PF3D7_1464500, and PF3D7_0510100, each mutated three times independently. Many of the highly overrepresented genes (7 of 14 genes mutated at least eight times independently), particularly drug targets, were also contained within CNVs. To identify genes likely to be involved in multidrug resistance, we plotted the likelihood of enrichment by chance against the number of associated compounds (**Figure 2.3B**). These data showed that genes such as *pfmdr1* and *pfap2-g* were mutated in selections with a wide variety of compounds. By contrast, targets such as *pfproRS* tended to be compound specific. This is not always the case, however, as some high-quality target genes, such as *pfcytb*, were identified in selections with a variety of scaffolds.

2.2.5 Network analysis reveals co-occurring mutations

To reveal gene interactions that arise in compound selections, we constructed a network with an edge between two genes if the combination of a specific mutation (SNV, indel, or CNV) in one gene and a specific mutation in the other gene was found in selections with different compounds (**Figure 2.3C**; see **methods**). This network was, in some cases, able to group genes by function. We identified well-known interactions, such as those between *pfat1* and *pfcarl*, or *pfacas* (PF3D7_0627800) and *pfacs11* (PF3D7_1238800)[81]. The data also showed that three of six compounds that gave rise to *pfcrt* SNVs also selected for amplifications of *pfabci3* (**Figure 2.3C**). There was a strong association between *pfap2-g* SNVs/indels and *pfmdr1* SNVs or CNVs. Of the

12 compounds yielding clones with *pfap2-g* mutations, seven also yielded *pfmdr1* mutations, and in five independent cases, both *pfap2-g* and *pfmdr1* mutations were found in the same clone. Interestingly, 11 of 14 different *pfap2-g* mutations were nonsense or frameshift, and only these loss-of-function *pfap2-g* mutations were found alongside *pfmdr1* amplification or missense SNVs in selections with the same compound. Furthermore, approximately 58% of genes on chromosome 12 with associations in the main network (7/12) were connected to both *pfmdr1* and *pfap2-g*. It may be that parasites cultured long term are likely to acquire loss-of-function mutations in *pfap2-g* because of pressure to eliminate genes causing gametocyte conversion in vitro, while *pfmdr1* point mutations arise infrequently. Another possibility is that *pfap2-g* regulates *pfmdr1*, and loss of *pfap2-g* function results in increased transcription of *pfmdr1*. Similarly, *pfapi-ap2* (PF3D7_0420300), a gene mutated in artemisinin selections[70], had associations with the conserved protein PF3D7_1324300. Another transcription factor, *pfapi-ap2* (PF3D7_0613800, ApiAP2_6 in **Figure 2.3C**), involved in the *Plasmodium* cell cycle and implicated in drug resistance evolution[82], had three associations with *pfpare* and ten with *pfmdr1*. This finding further supports the role of *pfapi-ap2* in contributing to multidrug resistance and culture adaptation in *Plasmodium*. *Pfap2-i* (chromosome 10) was part of a frequent amplification event. Although there are few reports of transcription factors playing a role in antimalarial drug resistance, this is common in other species such as *S. cerevisiae*[64].

2.2.6 Compound-selected mutations differ from naturally occurring variants

An important question in antimalarial drug development is whether natural parasite populations can readily attain resistance because of existing genetic variants and the plasticity of the *Plasmodium* genome. We compared our mutational set to nucleotide and amino acid

substitution rates in the Pf6 dataset[83], which includes sequences of 7,113 *P. falciparum* isolates obtained from 28 countries. SNVs that arose at least twice independently in our selections were rarely present in the Pf6 dataset (6 of 53). Genes with higher dN in the compound-selected set, measured as number of nonsynonymous SNVs normalized by nonsynonymous sites in the gene (see **methods**), tended to have low dN in Pf6. This suggests that in vitro selections typically introduced novel evolutionary pressures in genes that are highly conserved in the field, such as the proteasome β2 subunit (PF3D7_1328100), *pfdhfr*, and elongation factor 2 (PF3D7_1451100) (**Figure 2.3D**). We also show that multidrug resistance genes, such as *pfmdr1* or *pfcrt*, were less conserved in field isolates than canonical enzymatic drug targets. Our data show that 4,418 nuclear genes with SNVs (excluding singletons) in the Pf6 dataset had no nonsynonymous SNVs among the 724 evolved clones (excluding singletons). Among these are genes involved in antigenic variation and immune response, such as *pfcsp* (PF3D7_0304600), *pfama1* (PF3D7_1133400), *pfmsp2* (PF3D7_0206800), and *pfceltos* (PF3D7_1216600), which have nonsynonymous variation in the field but were not mutated in our compound selections[84, 85] (**Figure 2.3D**).

2.3 Discussion

Here, we present a comprehensive dataset of compound-selected resistance alleles for the *P. falciparum* malaria parasite. Similar to MalariaGEN's latest Pf7 dataset of field isolate sequences, we anticipate this resource and insights into shared characteristics of resistance-conferring alleles will be useful for several applications in the discovery and deployment of antimalarial drugs.

As we show, in vitro evolution of compound resistance typically gives rise to few mutations over the course of compound selection, compared to the thousands of genetic variants which

distinguish even slightly diverged isolates from the field. However, it remains the case that most SNV/indel mutations in our dataset likely do not drive compound resistance and instead are neutral mutations or improve fitness in in vitro culture. Moreover, not all genes enriched for in vitro evolved mutations are drivers of compound resistance; some may play roles in culture adaptation, while multigene families in non-core hypervariable regions were also frequently mutated. Further experimental work is needed to validate the roles of these overrepresented genes.

Importantly, our dataset can also serve to inform considerations of non-malarial resistance mechanisms. For example, the dataset can be a resource for identifying potential drug resistance alleles in microbial pathogens or cancer cells collected from patients treated with dihydroorotate dehydrogenase inhibitors, which frequently target the same ubiquinone-binding pocket in a range of organisms[86]. Our work suggests that drug resistance mechanisms are often conserved across species. As an example, cladosporin resistance has been associated with lysyl-tRNA synthetase mutations in yeast, and in *Plasmodium* cladosporin selected for amplifications of lysyl-tRNA synthetase[59].

There are limitations to using in vitro–selected mutations to inform drug development decisions. For instance, in vitro models of *Plasmodium* infection cannot recapitulate the selective pressures applied by the host immune system and other host-pathogen, host-vector, or host-xenobiotic interactions. Additionally, most laboratory strains of *P. falciparum* are not derived from recent parasite isolates. Rather, these parasites likely contain multiple adaptations to long-term culture, and their genetic backgrounds may not be fully representative of current natural parasite populations. Despite differences in selective pressure in the laboratory versus the field, insights from experimentally controlled compound selections are valuable.

Finally, our work reveals how to distinguish phenotype-driving variants from passengers, a key challenge of forward genetics approaches. This dataset and others could empower future machine learning-based approaches to estimate variant functionality in silico.

2.4 Materials and methods

2.4.1 Data acquisition

Sequencing information for parasite samples with evolved resistance to 25 antimalarial compounds were downloaded from the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) or acquired through direct correspondence with the senior author(s) of published work. These compounds include the gold-standard antimalarial drug artemisinin[70]; the benzoxaboroles AN13956, AN13762, AN10248[87]; the triazolopyrimidine analog series including DSM1[78], DSM265, and DSM267[88]; the imidazolopiperazine GNF179[89]; halofuginone, a synthetic derivative of the natural quinazolinone alkaloid febrifugine[90]; the boronate human proteasomal inhibitor bortezomib[91]; the pantothenamide CXP18.6-052[92]; the dihydroisoquinolones SJ733, SJ247, SJ311, and SJ279[93]; the 2,6-disubstituted quinoline-4-carboxamide DDD107498[94]; the antitubercular clinical candidate SQ109[95]; the peptide vinyl sulfones WLL-vs and WLW-vs[96]; the primary sulfonamide glycoside PS-3[97]; the bis-1,2,4-triazine compound MIPS-0004373[98]; MMV019313[99]; the trisubstituted imidazole MMV030084[100]; the 3-hydroxypropanamidine compound 22[101]; and the pyrazolopyrimidine sulfamate ML901[102].

2.4.2 In vitro evolution of compound-resistant parasites

P. falciparum parasites were cultured in RPMI1640 media supplemented with 0.5% AlbuMAX II, 4.3% human serum, 25 mM Hepes, 25 mM $NaHCO_3$, 0.36 mM hypoxanthine, and

100 ug/mL gentamicin. Cultures were maintained in leukocyte-depleted red blood cells (RBCs) at 2.5% hematocrit (HCT) and incubated at 37°C in an atmosphere of 3% O_2, 4% CO_2, and 93% N_2. In some cases only AlbuMAX II supplemented RPMI media without gentamicin and with 0.15 mM hypoxanthine, or only human serum supplemented RPMI media with gentamicin and with 0.15 mM hypoxanthine, was used in standard culture conditions. The human biological samples were sourced ethically and their research use was in accord with the terms of the informed consents. Resistant parasites were generated using either a high-pressure pulse, ramp-up, stepwise, or constant method of compound exposure as previously described[69]. Cultures were maintained under selection conditions until they demonstrated a reproducible IC_{50} fold-shift of at least 3×, at which point parasites were cloned in 96-well plates by limiting dilution[103].

Compound IC_{50} was assessed in dose-response format including using a SYBR Green-I based cell-proliferation assay as previously described[104]. Parasites were incubated for 72 h in 96-well or 384-well plates with exposure to the compound of interest in a 12-point dilution series. An artemisinin dilution series was conducted in parallel as a positive control. Following the incubation, parasites were lysed, and DNA was stained using SYBR Green fluorescence and was measured at 535 nm on an Envision plate reader (Perkin Elmer, Waltham, MA) or SpectraMax iD5 plate reader (Molecular Devices, San Jose, CA) after excitation at 485 nm. For some lines, *P. falciparum* growth inhibition was determined using a modified in vitro [^{3}H]-hypoxanthine incorporation method as previously described[77] or flow cytometry on an iQue Screener PLUS (Sartorius) with parasites stained with 1× SYBR Green and 200 nM MitoTracker Deep Red[98, 105]. Dose-response curves were fitted and log(IC_{50}) values calculated using Prism (GraphPad Prism, La Jolla, CA), or Excel and Grafit 5 software. IC_{50} fold-shift changes were calculated by comparing IC_{50} of the resistant clones to that of the corresponding compound-sensitive parent line.

2.4.3 Whole-genome sequencing analysis and annotation of variants

Raw sequencing reads were aligned to the *P. falciparum* 3D7 reference genome (PlasmoDB v13.0) and pre-processed following standard GATK version 3.5 protocols[68]. SNVs and indels were called with GATK HaplotypeCaller and filtered to retain high-quality variants. SNV calls were retained if they did not meet the following exclusion criteria: ReadPosRankSum > 8.0 or < −8.0, QUAL < 500, Quality by Depth (QD) < 2.0, Mapping Quality Rank Sum < −12.5, and filtered depth (DP) < 7. Indels were retained if they did not meet the following exclusion criteria: ReadPosRankSum < −20, QUAL < 500, QD < 2, and DP < 7. Allele fraction cutoffs for alternate alleles were ≥ 0.35 for bulk samples and ≥ 0.90 for clonal samples. Functional annotation of variants was carried out using SnpEff[106] with a custom database built from the 3D7 GFF from PlasmoDB (https://plasmodb.org/plasmo/app). Mutations that were considered background (native to the compound-sensitive parent) were removed, leaving a list of only high-quality mutations that had evolved over the course of the compound selection process.

CNVs were detected by calculating denoised $\log_2$ copy ratios across gene intervals through the GATK 4.0 CNV pipeline. We constructed two separate panels of normals for Dd2 and 3D7, using 30 independently sequenced parent clones of each genetic background. Read counts for each sample were calculated across a predefined gene interval list where intergenic regions and the highly variable *pfvar*, *pfrifin*, *pfstevor*, and *pfsurfin* genes[107] were removed. Following denoising against the strain-matched panel of normals, $\log_2$ copy ratios were calculated for each gene interval. CNVs were retained if at least 4 sequential genes showed a denoised $\log_2$ copy ratio of at least 0.4, indicating potential gene duplication.

To account for a subset of samples yielding false positives due to noisy or altered coverage profiles, we further filtered the candidate CNV list obtained from segmentation of genic copy ratios based on three additional criteria: statistical significance of difference in copy ratio vectors, discordant read-pair support for tandem duplications, and overlap with an independent CNV calling method, DELLY[108]. Based on manual inspection of select CNVs, we decided on a set of heuristics for filtering candidate CNVs to produce the final list. Samples were classified based on interquartile range (IQR) of core genome-wide gene copy ratios; stricter support thresholds were applied to those with high (> 0.3) or intermediate (0.12–0.3) IQR. Adjacent candidate CNVs were also combined. To optimize sensitivity and specificity on a manually validated test set, candidate CNVs were retained if they either had both low *P* value and sufficient overlap with DELLY calls, or both tandem duplication support (not near noncore regions) and some overlap with DELLY calls. The specific thresholds used in the filtering step are described in **Supplemental Table 2.3**.

2.4.4 Gene network construction

A subset of the variants was generated using SNV/indel (**Supplemental Table 2.2**) and CNV (**Supplemental Table 2.3**) data filtered for core genes and mutation types, i.e. missense, disruptive inframe insertion or deletion (indel), frameshift, start loss, stop gained or lost, splice region variant and combinations of these. The list was processed to create pairs of compound-gene(s) and compound-gene-variant type. The network was generated using these pairs, having genes as nodes, and edges where two genes have at least once instance of a pair of compounds yielding resistant samples with the same SNV/indel (gene 1)–SNV/indel (gene 2) or SNV/indel (gene 1)–CNV (gene 2) variant pair. The network was visualized using Cytoscape v.3.9.1 organic layout[39]. The node color represents the score calculated based on the variant type adjusting for

nonsense (light blue, lower score) to missense (dark blue, higher score); the total of variants is shown by the node's size adjusting for CNV (smaller) vs SNV/indel (bigger). Edge intensities show the total number of mutations combined by the compound pair from light grey (less mutated) to blue (highly mutated). Highly variable multigene families (*pfvar*, *pfrifin*, *pfstevor*, and *pfsurfin*) were removed from the final network.

2.5 Acknowledgements

Chapter 2 is a partial, modified reprint of the material as it appears in Science, 2024. Madeline R. Luth, Karla P. Godinez-Macias, Daisy Chen, John Okombo, Vandana Thathy, Xiu Cheng, Sindhu Daggupati, Heledd Davies, Satish K. Dhingra, Jan M. Economy, Rebecca C.S. Edgar, Maria G. Gomez-Lorenzo, Eva S. Istvan, Juan Carlos Jado, Gregory M. LaMonte, Bruno Melillo, Sachel Mok, Sunil K. Narwal, Tolla Ndiaye, Sabine Ottilie, Sara Palomo Diaz, Heekuk Park, Stella Peña, Frances Rocamora, Tomoyo Sakata-Kato, Jennifer L. Small-Saunders, Robert L. Summers, Patrick K. Tumwebaze, Manu Vanaerschot, Guoqin Xia, Tomas Yeo, Ashley You, Francisco-Javier Gamo, Daniel E. Goldberg, Marcus C.S. Lee, Case W. McNamara, Daouda Ndiaye, Philip J. Rosenthal, Stuart L. Schreiber, Gloria Serra, Jair L. De Siqueira-Neto, Tina S. Skinner-Adams, Anne-Catrin Uhlemann, Nobutaka Kato, Amanda K. Lukens, Dyann F. Wirth, David A. Fidock, Elizabeth A. Winzeler. "Systematic in vitro evolution in *Plasmodium falciparum* reveals key determinants of drug resistance". The dissertation author was a co-first author of this paper.

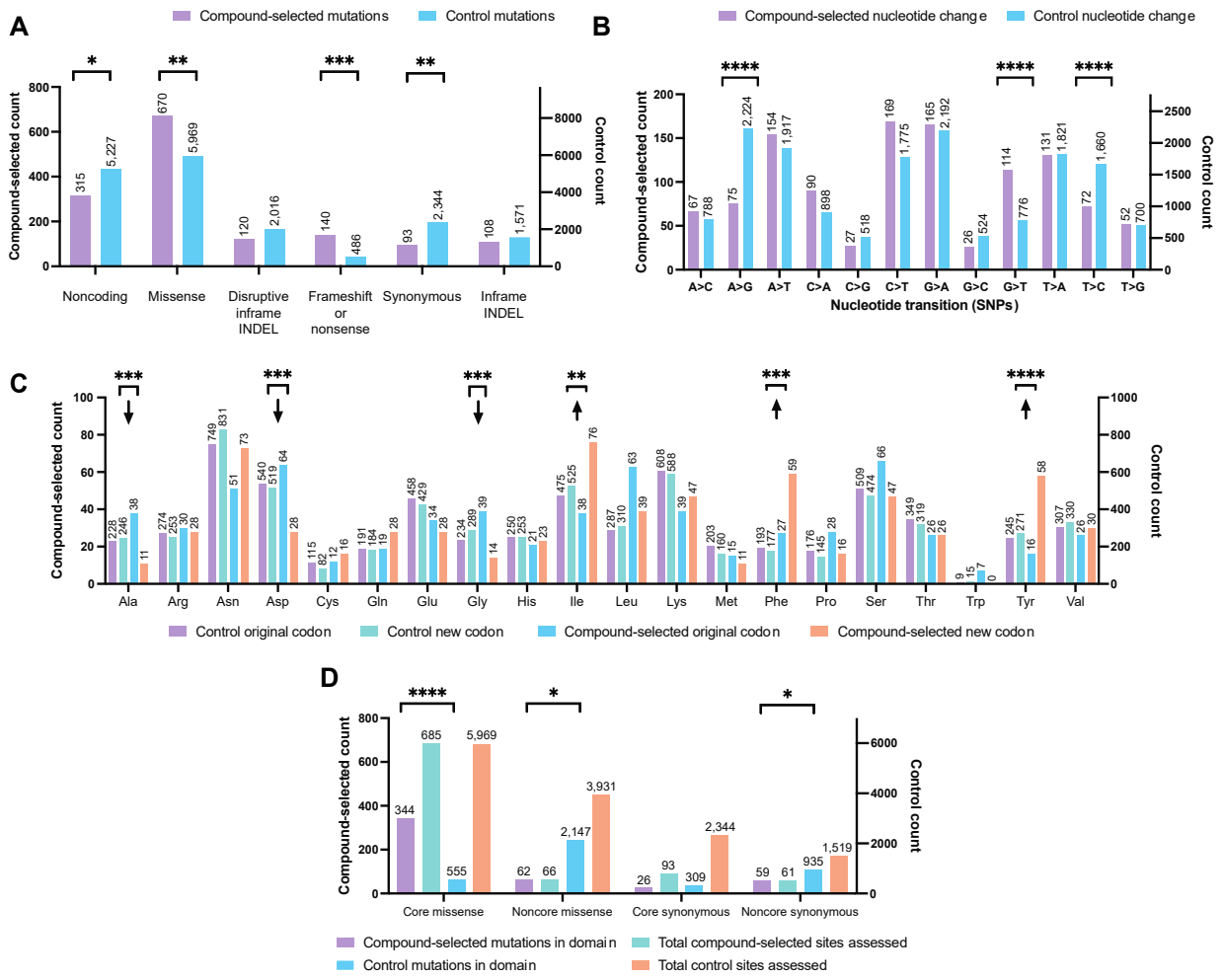


Figure 2.1 Compound-selected single nucleotide variants (SNVs) and indels have different characteristics compared to naturally occurring variants

(A-D) Comparison between the set of 1,448 compound-selected variants (SNVs and indels) in the core genome from this study (left axis) and a set of 17,613 control variants obtained by aligning Dd2 to the 3D7 reference genome. Axes have been normalized to display differences in proportions of different categories for compound-selected (left) and control (right) variants. **(B)** Reference base and alternate base for 1,141 unique core SNVs (coding and noncoding) compared to 15,793 control variants. **(C)** Analysis of amino acid changes for 640 core missense and 6,600 core missense control SNVs with arrows showing relative increases or decreases. **(D)** 905 evolved and 13,763 control protein coding variants were analyzed for location within a defined InterPro domain obtained from PlasmoDB v61. In all **** indicates $P < 0.0001$; *** < 0.001 and ** < 0.01 calculated using a Chi-square test.

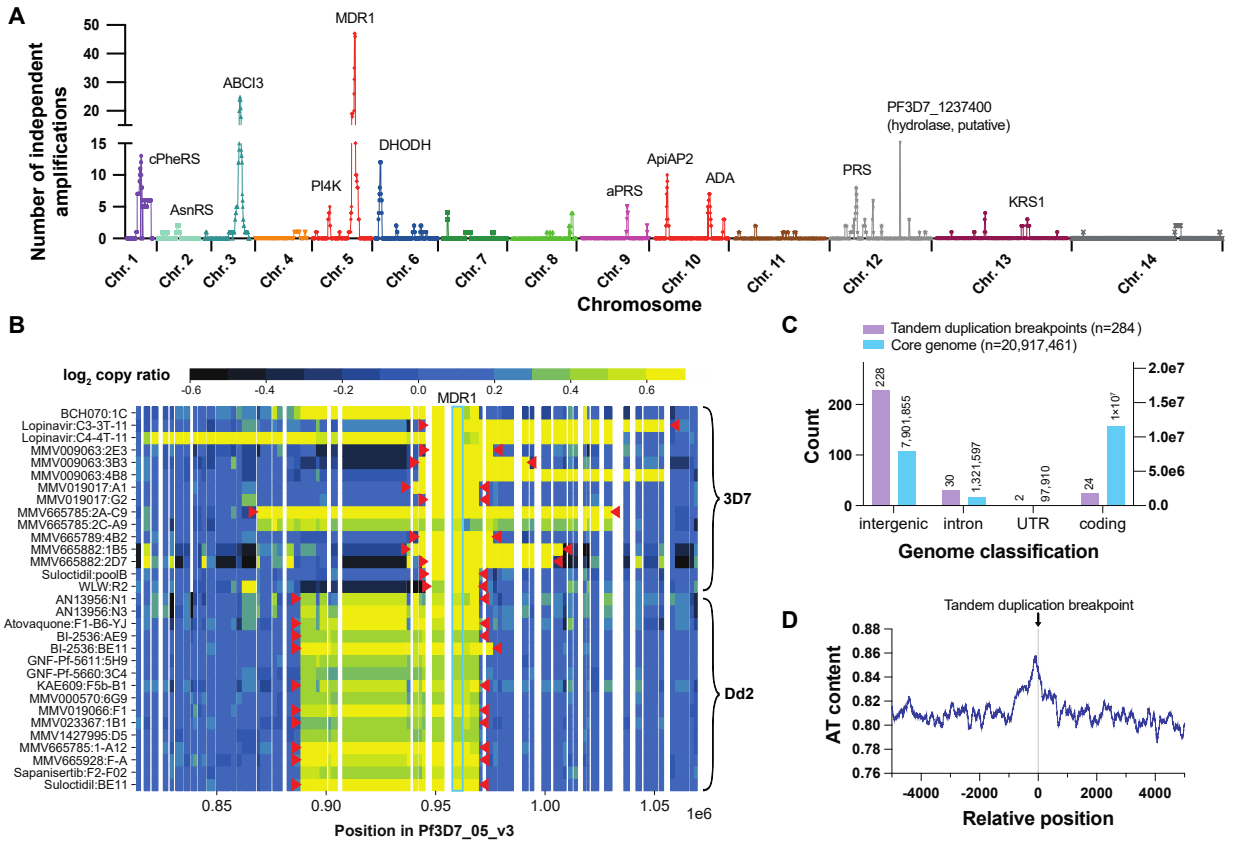


Figure 2.2 Copy number variants (CNVs) are a frequent driver of antimalarial compound resistance

(A) Manhattan plot showing the number of times each gene was amplified as part of an independent CNV in our dataset. Genes that are likely drivers of the selective advantage conferred by CNVs are annotated, including compound targets and multidrug resistance genes. **(B)** Heatmap visualization of amplifications containing *pfmdr1* for evolved clones (*pfmdr1* CNVs that independently involved in the same compound selection were omitted for brevity). Denoised $\log_2$ gene copy ratios are plotted for each clone; a contiguous segment of high $\log_2$ copy ratio (yellow) suggests an amplification including those genes. For CNVs identified as tandem duplications, more precise CNV boundaries are indicated by red markers. **(C)** Comparison of genomic classification distribution between 284 independent tandem duplication CNV breakpoints and all sites in the core genome. **(D)** AT content around tandem duplication breakpoints, averaged over the 284 independent tandem duplications. AT content for each relative position was computed as the proportion of bases that are A or T over all sequences beginning at that position with a sliding window of five bp. The plot is shown as a running average with a window of 100 bp.

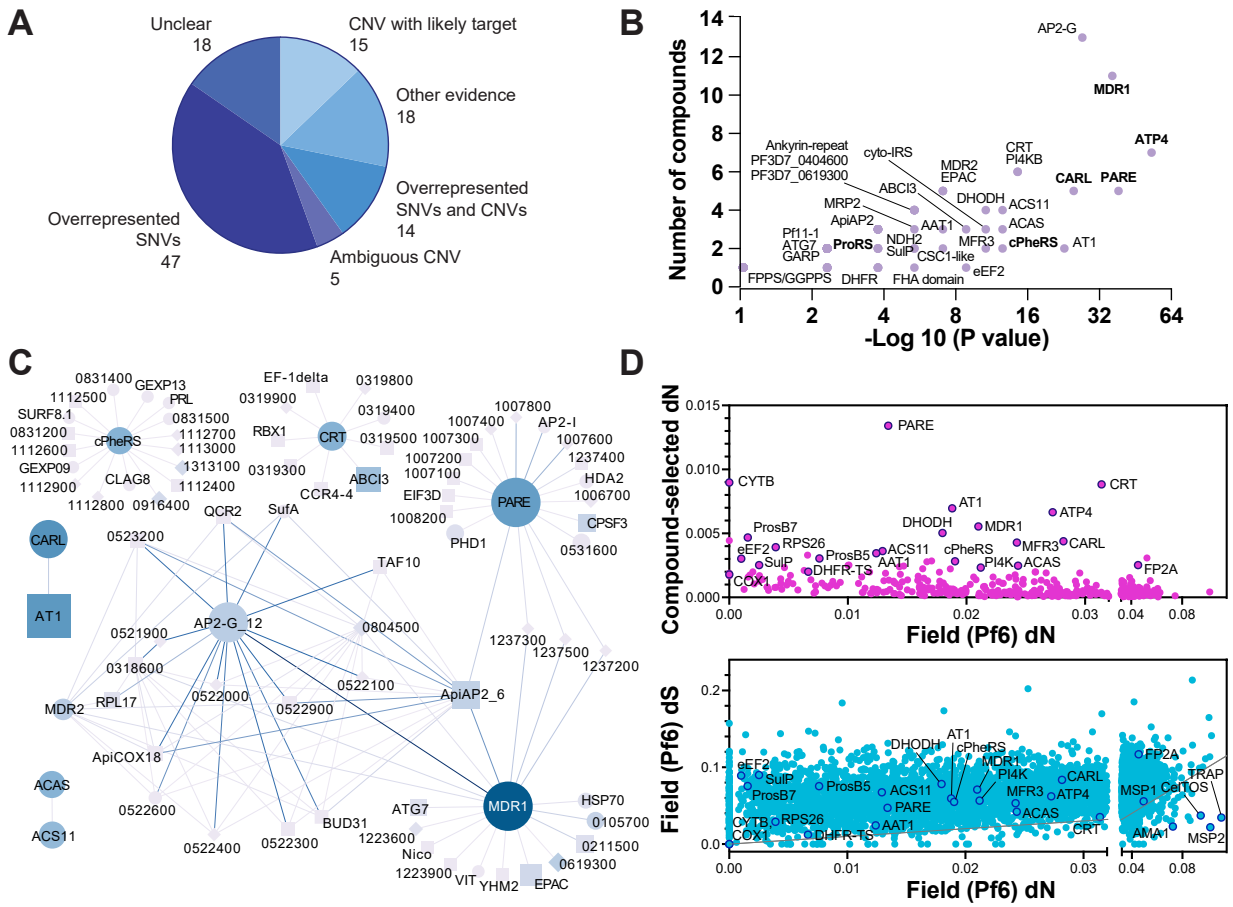


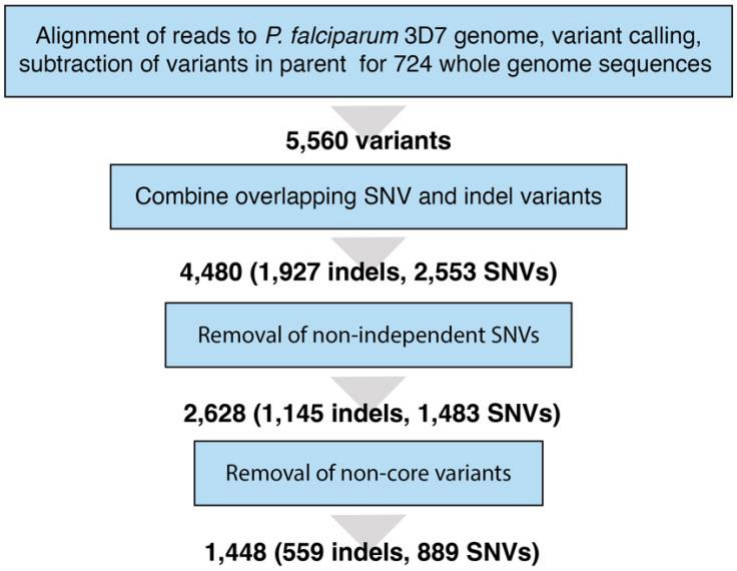
Figure 2.3 Enriched genes mutated in compound resistance selections

(A) Classification of evidence types supporting assignment of target or resistance mediator for 118 compounds (**Supplemental Table 2.4**). "Other evidence" indicates the target/resistance gene was identified but its representation did not reach statistical significance (hypergeometric test); other experimental evidence was required for confirmation. "Ambiguous CNVs" were cases in which a CNV was found in a majority of selected clones but a clear target or resistance gene was not identified. **(B)** Enrichment *P* values (hypergeometric mean function) for genes that were recurrently mutated in selections for a given compound (**Supplemental Table 2.1**). **(C)** Condensed version of a network linking genes for which the same pair of mutations arose in distinct compounds, filtered to show gene pairs with at least ten shared variant pair-compound pair occurrences. In cases of more than one variant pair-compound pair existing between two genes, this multiplicity was encoded as edge weight. All seven digit gene IDs have PF3D7_ removed. Circles represent known proteins; diamonds, conserved proteins; squares, putative proteins. Node size is proportional to node degree and edge color maps to occurrence of disruptive mutations from low (light blue) to high (dark blue) after adjustment for non-missense mutations. **(D)** dN for 449 genes that had more than one non-singleton SNV in the compound-selected dataset plotted against dN among Pf6 samples (top) and dS vs. dN of 4,938 core genes in the Pf6 dataset of worldwide variation (bottom).

Table 2.1 Compound library used to test blood-stage parasites resistance.

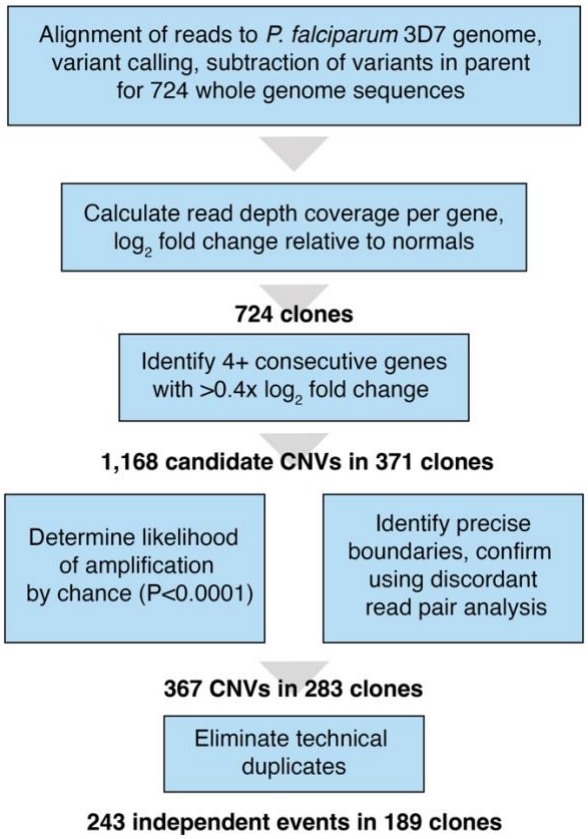
Antimalarial compounds used in this study, including reference PubMed reference if applicable.

Compound	Compound	Compound	Compound
3-Hydroxy-propana-midine 22 (33666415)	GNF452 (24589256)	MMV020746 (29326268)	MMV666080 (27301419)
ACT-451840 (27073104)	GNF707 (24589256)	MMV023367 (29326268)	MMV667491 (34233174)
AN10248 (31992618)	Halofuginone (25995223)	MMV023388 (28674055)	MMV668311 (34233174)
AN13762 (31992618)	KAE609 (23414762)	MMV023949 (28674055)	MMV668399 (29326268)
AN13956 (31992618)	KAE678 (25322084)	MMV024038 (27301419)	MMV673482 (27602946)
Antimycin A	KAI407 (24284631)	MMV024101 (33334080)	MMV675939 (34233174)
Artemisinin (29538461)	Lopinavir (15980379)	MMV024114 (29326268)	MMV676881 (33932049)
Atovaquone (23408914)	MIPS-0004373 (34460304)	MMV026596 (29326268)	MMV688533 (34233174)
BCH070 (31616928)	ML901 (35653481)	MMV027634 (29326268)	MMV907364
BI-2536 (36919909)	MMV000570 (27467575)	MMV028038 (29326268)	MMV973707 (33929818)
BMS983970	MMV006767 (29326268)	MMV029272 (29326268)	MMV982556 (30655961)
Bortezomib (30373366)	MMV006901 (28674055)	MMV030084 (32359426)	NCP26 (36008486)
BRD1095 (27602946)	MMV007224 (29326268)	MMV084864 (34233174)	NEU1029 (36942408)
BRD3444 (27602946)	MMV007564 (27933786)	MMV084978 (34348113)	OSM-S-106 (37546892)
BRD9685	MMV008149 (29326268)	MMV085203 (33715347)	Pepstatin n-butyl ester (28106035)
Cladosporin (22704625)	MMV008434	MMV1081413 (36888694)	Primaquine (29326268)
CXP18.6-052 (31534021)	MMV009063 (29326268)	MMV1091186 (36888694)	Proguanil (31069275)
DDD01061024 (32078764)	MMV009108 (29326268)	MMV1427995 (30523084)	PS-3 (33293381)
DDD107498 (26085270)	MMV010545 (28674055)	MMV1432711 (30523084)	Sapanisertib (36260689)
Desoxyepothilone B	MMV011438 (29326268)	MMV1454442 (30523084)	SJ247 (25453091)
DSM1 (18522386)	MMV011895 (29326268)	MMV1490397 (30523084)	SJ279 (25453091)
DSM265 (31801884)	MMV019017 (29326268)	MMV1490405 (30523084)	SJ311 (25453091)
DSM267 (31801884)	MMV019066 (29326268)	MMV407834 (28674055)	SJ733 (25453091)
GNF-Pf-3600 (33715347)	MMV019313 (29276048)	MMV665789 (29326268)	SQ109 (33431834)
GNF-Pf-4492 (25322084)	MMV019662 (29326268)	MMV665794 (37244916)	Suloctidil
GNF-Pf-5310/MMV665785 (28808258)	MMV019719 (29326268)	MMV665852 (29326268)	TCMDC-124553 (36888694)
GNF-Pf-5611 (18579783)	MMV019721 (34348113)	MMV665882 (29326268)	TCMDC-125334 (29336544)
GNF-Pf-5660 (28808258)	MMV020623 (28674055)	MMV665924 (29326268)	Verapamil
GNF-Pf-5668 (28808258)	MMV020746 (29326268)	MMV665928 (27301419)	WLL (31170268)
GNF179 (27381290)	MMV023367 (29326268)	MMV665939 (29326268)	WLW (31170268)



Supplemental Figure 2.1 Indels and SNV detection workflow.

Pipeline for variant detection and filtering to identify variants among the 724 clones analyzed. These variants were used as part of **Figure 2.1**.



Supplemental Figure 2.2 CNV detection workflow.

Pipeline for CNV detection and filtering to identify amplification CNVs among the 724 clones analyzed. These CNVs were used as part of **Figure 2.2**.

Chapter 3

Revisiting the *Plasmodium falciparum* druggable genome using predicted structures and data mining

3.1 Introduction

Over the past decade, phenotypic screening has gained popularity since large, diverse compound libraries can be tested in a high-throughput fashion without requiring a priori knowledge of targets or mechanisms of action[67]. After triage, a subset of screening hits is typically subjected to target identification, facilitating target-based medicinal chemistry programs. In malaria parasites, this approach has revealed new targets[66, 67], including P-type cation translocating ATPase 4 (ATP4)[109], acetyl-CoA synthetase (ACAS)[55], and several aminoacyl-tRNA synthetases[110-113]. Nonetheless, novel targets are urgently needed to develop drugs that differ from existing antimalarials in mechanism of action and resistance liability.

Compound-dependent target discovery faces several limitations. Limited chemical matter and follow-up on only the most potent phenotypic hits pose restrictions on the targeted biology space. As a result, studies tend to reidentify targets such as *Pf*DHODH and *Pf*ATP4[67]. In addition, target identification is an arduous process, especially when targets lack known ortholog inhibitors in other species. Alternatively, in silico approaches can identify proteins amenable to target-based drug discovery[67, 114]. For malaria parasites, key characteristics of a candidate target are its "druggability" (i.e., ability to be modulated by a drug-like ligand), and its essentiality to parasite survival in the lifecycle stage of interest. Although the druggable genome of the deadliest human malaria species, *Plasmodium falciparum*, has been explored via in silico methods[69, 115], these studies have relied on homology to known targets and drug–gene interactions to predict druggability[115, 116]. Homology-based protein modelling tools like SWISS-MODEL[117] and molecular docking tools like AutoDock Vina[118], while useful for virtual screening[67], are less appropriate for assessing proteome-wide druggability. With the advent of artificial intelligence models for predicting protein structure, such as AlphaFold[119] and ESMFold[120], along with ligand-binding prediction tools like AlphaFill[121], we are now able to comprehensively assess the whole

genome for essential proteins with evidence of small molecule binding. This approach may identify drug targets overlooked by compound-dependent discovery efforts.

Leveraging the collective expertise of the Malaria Drug Accelerator (MalDA)[66] consortium—a partnership between academia and industry aiming to accelerate antimalarial drug discovery— we systematically identified and ranked a list of "druggable target" candidates from the entire *P. falciparum* genome that could progress into drug discovery. The list was determined by identifying genes with evidence of both protein binding to small molecules[121-123] and essentiality in the parasite asexual blood-stage (ABS)[124-126]. Their viability as drug target candidates was further assessed based on common characteristics of known drug targets using available literature and data[67, 127-130]. In addition to highlighting promising blood-stage antimalarial targets, we provide an in-depth annotation resource that can facilitate future target validation and lead optimization efforts. By predicting ligand-target interactions from AlphaFill[121], BindingDB[122], and BRENDA[123], we uncovered some small molecule tool compounds for further studies or as starting points for structure activity relationship (SAR) design. The framework used in this study to integrate and evaluate druggability information may be applied to genome-wide in silico target discovery for other pathogenic organisms and diseases.

3.2 Results

3.2.1 Defining 1,660 *P. falciparum* genes with evidence of small molecule binding

From 5,318 protein-coding genes in the *P. falciparum* 3D7 genome (PlasmoDB[131] release 66), we identified a set of proteins that are "ligandable" and, therefore, potentially druggable. Proteins were evaluated by integrating data from predictions of ligands based on similarity to existing co-crystallized protein structures using AlphaFill[121], orthology or sequence similarity to

proteins with experimentally determined protein–ligand–binding affinities in BindingDB[122], and curated enzyme-inhibitor interactions in BRENDA[123] (**Figure 3.1A**, **Supplemental Table 3.1** and **3.2**).

Of the 5,099 *P. falciparum* 3D7 genes with an AlphaFold protein model, 2,771 had at least one AlphaFill "hit", i.e. sufficient local sequence homology to a protein in the PDB-REDO databank[132] associated with ligand(s), referred to as potential "transplants". We restricted our attention to 1,233 proteins that had at least one confident AlphaFill transplant with global RMSD (root-mean-square-deviation, a measure of structural similarity) < 10 and local RMSD < 4, thresholds informed by empirical observation. Precipitants commonly used in protein crystallization and small ligands (< 10 atoms) were ignored, as they are unlikely to be drug-like (see **methods**). To broaden druggability evidence for *P. falciparum* enzymes and overcome the fact that many *P. falciparum* proteins are not orthologous to crystallized proteins, we incorporated information on inhibitors linked to EC (Enzyme Commission) number classes in the BRENDA database[123]. This yielded 321 additional proteins lacking confident AlphaFill predictions (**Supplemental Table 3.1**). We further augmented our ligandable set using 6,202 targets (UniProt IDs) from BindingDB[122], a curated database of experimentally determined protein–ligand-binding affinities. Of 5,318 *P. falciparum* proteins, 581 were orthologous to at least one of the 6,202 BindingDB targets based on OrthoMCL[127], OMA[129], HOGENOM[133], or OrthoDB[128] phylogenomic databases, or based on BLAST[134] hits (*E* value < 1) to the OrthoMCL full protein database (**Supplemental Table 3.2**).

Altogether, we found 1,660 unique proteins with at least one source of small molecule-binding evidence. Many (n = 817) were identified by only one source, demonstrating the importance of multiple evidence sources to reduce false negatives. The set may include a few false

positives; one possible example is the apical membrane antigen 1 (AMA1, PF3D7_1133400), an essential vaccine candidate lacking evidence of classical druggability. In crystallography studies[135], Akter et al. used a peptide probe with the spin label MTSL, also known as (1-Oxyl-2,2,5,5-tetramethylpyrroline-3-methyl)-methanethiosulfonate, which was identified as an AlphaFill hit and is of similar size to small molecule inhibitors[135]. More sophisticated approaches than molecular weight filtering may be needed to reduce false positives, highlighting the need for expert review, as described below.

To assess our approach, we examined a set of 43 known *P. falciparum* antimalarial targets that have some level of clinical, in vivo, or in vitro validation[67]. We found that, except for NCR1 (PF3D7_0107500), all were supported by at least one source of binding evidence (**Figure 3.1B**). Twenty-six of the 43 validated targets were well-known enzyme targets such as DHFR-TS (PF3D7_0417200) and DHODH (PF3D7_0603300) that had AlphaFill hit(s), ortholog(s) in BindingDB and known enzyme class inhibitors from BRENDA. In five cases (eEF2, elongation factor 2; CPSF3, cleavage and polyadenylation specificity factor subunit 3; FNT, formate-nitrite transporter FNT; PF3D7_1038900, a monoacylglycerol lipase-like esterase; and MQO, malate:quinone oxidoreductase), a single source of binding evidence rescued the validated target.

3.2.2 Defining 1,929 *P. falciparum* genes with evidence of blood-stage essentiality

To assess which of the 5,318 *P. falciparum* protein-coding genes are required for asexual parasite growth, we incorporated essentiality data for *P. falciparum* and the rodent malaria species *P. berghei*, reasoning that orthologous genes could provide additional information on essentiality. We focused on the asexual blood-stage due to its role in the manifestation of clinical symptoms as well as completeness of available essentiality screens. In the *P. falciparum* screen by Zhang et

al.[124], 3,271 proteins were labeled essential for in vitro ABS growth based on genome-wide transposon mutagenesis. Among 2,383 *falciparum* orthologs of *berghei* genes tested with gene disruption vectors in the PlasmoGEM dataset[125], 1,145 were essential in the ABS, and the RMgmDB dataset[126] indicated a change in phenotype upon gene modification for 1,319 of 1,609 *P. falciparum* genes whose *berghei* orthologs were tested[125, 126].

Reasoning that ambiguous essentiality data should not preclude proteins from consideration as targets, we created a categorization scheme for strength of essentiality evidence (**Supplemental Figure 3.1**). Categories were defined as "clear support", "unclear support", "unsupported", or "no data" for essentiality in the asexual blood-stage. For "clear support", all available evidence sources must confidently label the protein as essential; if either Zhang et al. or PlasmoGEM confidently labeled the protein as nonessential, it was considered "unsupported", while "unclear support" describes all other proteins with data from at least one source.

In total, 1,929 *P. falciparum* proteins were classified as having clear essentiality support, 1,008 with unclear support, 2,326 with support for non-essentiality, and 55 with no data (**Figure 3.1C**). Surprisingly, this classification scheme categorized five of the 43 validated targets from Siqueira-Neto et al.[67] (MQO, PDEdelta, PNP, PF3D7_1038900, and PMX) under unclear support (**Figure 3.1D**). In all but one case, either the *P. falciparum* or *berghei* datasets suggest the protein is essential while at least one source is contradictory. The exception was PDEδ (PF3D7_1470500, cGMP-specific 3',4'-cyclic phosphodiesterase δ), which regulates erythrocyte deformability[136] and is not a blood-stage target but rather the target of tadalafil in mature gametocyte stages. While these results show that available data are sometimes inconsistent and can only partially inform *Plasmodium* gene essentiality, by combining multiple evidence sources, we increased our

confidence that proteins categorized as "clear support" are essential and thus potential antimalarial targets.

3.2.3 867 *P. falciparum* proteins have evidence of binding and blood-stage essentiality

To define an initial list of candidate targets, we took the intersection of the 1,660 proteins with small molecule binding evidence and the 2,992 proteins not categorized as "unsupported" (**Figure 3.1C**). This yielded 867 candidate targets after filtering 19 genes in hypervariable noncore regions (**Supplemental Table 3.3**). Noncore genes, encompassing *var*, *rifin* and *stevor* multigene families and other genes in highly recombinogenic subtelomeres[137, 138], were not considered as their variability and redundancy make them poor targets even where deemed essential (e.g., PF3D7_0101600, a rifin with mutagenesis index score of 0.199[124]). The 867 candidate targets were distributed throughout the genome with no apparent propensity for specific chromosomes. Most (n = 651) of the 867 candidate targets were supported by AlphaFill binding evidence, 336 were orthologs of or had BLAST matches to validated targets in BindingDB, and 457 were supported by BRENDA enzymatic data. Among 857 candidate targets present in the Zhang et al. *P. falciparum* dataset, 850 were labeled as essential, in contrast to 2,421 of 4,396 non-candidate proteins (**Figure 3.2A**).

Attractively, 577 of the 867 candidate targets were confidently essential in both *P. falciparum* parasites and the PlasmoGEM (n = 452) or RMgmDB (n = 162) *P. berghei* datasets (**Figure 3.2A**). This suggests potential as therapeutic targets for multiple *Plasmodium* species, which is important given that antimalarial drugs will need to act against *P. vivax* as well. Among these 577 candidates, we observed known targets (n = 7) with validated clinical inhibitors, like eEF2 and PI4K (phosphatidylinositol 4-kinase), and clinically unexplored targets such as BDP1

(PF3D7_1033700, bromodomain protein 1). A small number of candidate proteins (n = 14) appeared to be confidently essential in *P. falciparum* but not *P. berghei*, including two acyl-CoA synthetases (PF3D7_0301000, PF3D7_0525100) and two serine/threonine FIKK kinases (PF3D7_0301200, PF3D7_0902400).

3.2.4 Building an annotation resource using scientific evidence to prioritize candidate targets

To evaluate the 867 candidate targets, we compiled additional annotations for all *P. falciparum* protein-coding genes (**Supplemental Table 3.2**). We included information on genomic features and genetic variation (PlasmoDB[131], NCBI[139] and MalariaGEN[60]), protein features and structures (Protein Data Bank[130] (PDB) and PlasmoDB), expression across the parasite lifecycle stages (Malaria Cell Atlas[140] and Le Roch et al.[141]), literature references (NCBI and PubMed), and similarity to human orthologs (**Figure 3.2A**, and **Supplemental Table 3.3**) (see **methods** for more details). We reasoned that in addition to druggability evidence, these annotations would allow us to prioritize proteins that merit further structural/functional characterization and target-based screening.

Only 286 of the 5,318 *P. falciparum* proteins have an experimentally determined structure in PDB database. Of these, 112 were in our list of 867 candidate targets, indicating prior characterization of many candidate targets and highlighting those amenable to structure-based drug design (**Figure 3.2A**). Examples of candidate crystal structures include ferredoxin-NADP reductase (FNR)[142] and aspartate carbamoyltransferase (ATCase) in complex with a recently discovered small molecule allosteric inhibitor[143]. We also observed that 2,006 protein-coding genes have human orthologs based on OrthoMCL. To estimate the structural similarity of *P.*

falciparum proteins to their human counterparts, we performed pairwise comparisons of their AlphaFold models using TM-align[144]. This allowed us to identify the most similar human ortholog for 1,972 *P. falciparum* proteins (AlphaFold structures were not available in 34 cases). On average, orthologs showed 33% sequence identity for local alignments that were, on average, 244 amino acids long. A human ortholog was not reported for 217 candidates, which may include promising targets involved in parasite-specific essential biology.

To characterize the biological functions of proteins in the candidate list, we performed Gene Ontology (GO) term enrichment analysis (**Figure 3.2B**). Across the 867 candidates, 857 had at least one associated GO term. The most highly enriched terms (tree depth > 2) were related to small molecule binding, in particular nucleotide binding (n = 299, P = 8.2x10^{-96}, Bonferroni corrected to reduce false positives). The cellular component term "intracellular organelle" was also highly enriched (n = 659, P = 1.2x10^{-71}). Other overrepresented GO terms among candidate targets include ATP binding (n = 213, P = 2.3x10^{-62}), pyrophosphatase activity (n = 119, P = 1.4x10^{-46}), and more. These results highlight differences in cellular function and localization between proteins predicted to be essential and ligandable and those that are not.

To facilitate candidate target evaluation, we gathered existing evidence by querying literature repositories using PlasmoDB and Entrez gene identifiers (**Figure 3.2C**). Through this approach, we rescued evidence predating the standardization of *Plasmodium* gene nomenclature; for example, four references for *pfhsp101* (PF3D7_1116800) were recovered. Overall, we found literature references for 4,956 genes, with a median of four references per gene among candidate targets and two references per gene among non-candidates (**Figure 3.2C**). Unsurprisingly, well-studied genes such as the multidrug resistance genes *pfcrt* (70 references) and *pfmdr1* (66

references), and vaccine targets such as *pfmsp1* (63 references) and *pfama1* (56 references), had the most references.

Finally, to identify candidates with multistage activity, we examined gene expression evidence across the parasite lifecycle using the Le Roch et al. microarray dataset[141]. This dataset is particularly valuable as it includes probabilities of detection above background. As expected, 85% (n = 737) of the 867 candidate targets were strongly supported by expression in at least one ABS substage, contrary to 64.5% of all *P. falciparum* genes (**Figure 3.2D**). Of these 737 candidates, 577 also showed strong evidence of expression in the sexual (gametocyte) or mosquito (sporozoite) stages. The remaining 130 candidates were either not measured (n = 36), had unclear expression (n = 57), or were clearly not expressed across ABS substages (n = 37). Around half of these 37 candidate genes also appeared to be minimally expressed according to ABS scRNA-seq data from the Malaria Cell Atlas study[140], while the other half either contradicted Malaria Cell Atlas expression levels or had dubious evidence of essentiality. In the latter case, many were small proteins (around 100 amino acids long), which have a lower probability of being detected with RNA-seq or tiling microarrays. It is possible that some proteins such as RPUSP (RNA pseudouridylate synthase) and YTH1 (YTH domain-containing protein) are essential despite low expression levels, which could be advantageous for an antimalarial target[145].

3.2.5 Scoring 540 novel candidate targets with strong evidence of essentiality

We next sought to narrow down the candidate targets to those that have strong evidence of essentiality and are relatively novel (limited prior characterization, especially as an antimalarial target). Starting from 587 candidate targets classified as “clear support” for ABS essentiality, we filtered well-known antimalarial targets, such as DHODH and DHFR-TS, validated targets, and

target classes currently being pursued by MalDA or others, such as aminoacyl tRNA synthetases[66]. This resulted in 540 understudied (novel) candidate targets with binding evidence suggesting they are likely to disrupt parasite growth and survival upon perturbation (**Figure 3.1C**).

Taking advantage of the data compendium, we created a rubric (**methods**) to manually score the 540 candidate targets based on their potential for antimalarial target-based drug discovery (**Figure 3.3A**, and **Supplemental Table 3.4**). Each target was scored on ten weighted categories, totaling to 100 points. The rubric considered the quantity and quality of compiled evidence, the readiness of functional or binding assay development, evidence of druggability and novelty across scientific literature; points were deducted for weak, missing, or contradictory evidence. Some target candidates were deprioritized by reviewers if concerning characteristics regarding their feasibility were observed. For example, reviewers deprioritized CK2α and FKBP35 due to lack of effect on asexual growth from conditional knockout studies[146, 147].

Under this rubric, scores for the 540 novel candidate targets ranged from 6 to 96 points, with an average score of 48.64 (**Supplemental Table 3.4**). Candidates with prior characterization tended to receive high scores, while lower scores (≤ 45) were assigned in the absence of recombinant protein expression, biochemical assays, and structural or druggability information. Among 255 low-scoring candidates were subunits of protein complexes, challenging enzyme classes like GTPases, and unsuccessful pre-clinical targets in other organisms. For example, RRP45 (PF3D7_1364500), an RNA exosome complex component, scored 36 points due to lack of recombinant protein production, lack of a biochemical assay and limited protein structure and tool compound information. Nevertheless, targets with limited prior work scored high in novelty according to our metric (**Supplemental Table 3.4**).

Our scoring also revealed attractive high-scoring candidates. One example is TopoI (PF3D7_0510500, topoisomerase I; 80 points), an enzyme involved in DNA replication, transcription, and repair (**Figure 3.3B**). A bacterial TopoI inhibitor[148] is known, suggesting that the *Plasmodium* enzyme could be selectively targeted. Although our attention was drawn to high-scoring genes, those with lower scores could still be potential drug targets. Such candidates, including the NAD kinase PF3D7_0913300 and proteins that lack human orthologs but are conserved within natural parasite populations like PF3D7_1446800 (heme detoxification protein), will require substantial additional research to confirm their viability as antimalarial targets.

3.2.6 Secondary scoring of 67 high-ranking candidate targets

Although candidate targets were scored with a predefined rubric, scores were manually determined and could thus vary among reviewers. For example, a reviewer may give a higher score if there is an enzymatic assay specifically for the enzyme under review, whereas another reviewer could give the same score if an enzymatic assay is available for the enzyme class. Therefore, to increase confidence in the scores, we conducted a second round of scoring using the same rubric (**methods**), for 67 high-ranking candidates (**Figure 3.3C**, **Table 3.1**, and **Supplemental Table 3.4**). These 67 candidates were chosen by selecting up to two of the highest-scored proteins recommended by each of the initial reviewers with a minimum first score of 50.

Secondary scoring for the 67 candidates averaged 69.22 points, slightly lower than the first round (73.55 points). Six candidates showed a difference of more than 20 points (**Figure 3.3C**). One example, ribosome biogenesis GTPase A (RbgA), decreased from 81 to 54 points, as the second reviewer placed greater emphasis on the lack of a tool compound and the fact that recombinant protein was only expressed in bacteria. On the other hand, seven genes received the

same score from independent reviewers, including ATCase and FNR (**Figure 3.3C**), supporting the rubric's utility in assessing candidate targets.

3.2.7 In-depth consideration of 27 prioritized candidates reveals targets poised for drug discovery

From the 67 high-ranking candidate targets with secondary scores, 27 were selected for in-depth consideration by a panel of MalDA experts by once again selecting the top 1–2 candidates recommended by each secondary reviewer (**Table 3.1**). Among these targets, two groups of apicoplast targets were highlighted by different reviewers: caseinolytic protease ATPases (ClpQ, ClpS, ClpY, ClpP, ClpB1) that play important roles in protein homeostasis, and methylerythritol phosphate enzymes (IspD, IspE, IspF) involved in the isoprenoid biosynthesis pathway. We also observed that seven of these 27 prioritized targets lack a human ortholog, suggesting their potential as highly selective targets. This exercise also highlighted five attractive targets: ATCase, TopoI, GyrB (DNA gyrase subunit B), GluPho, and BDP1. These five targets showed the least concerns for progressing in drug discovery efforts according to our rubric, with all but BDP1 having previously demonstrated small molecule inhibitors[149-152].

3.3 Discussion

Here, we present a systematic data compendium of the *Plasmodium* genome focused on druggability potential as well as an updated set of targets that can readily progress into drug discovery programs. To assess druggability evidence, we leveraged the AlphaFill database of predicted ligand "transplants" based on homology of AlphaFold structures to all structures in the PDB-REDO databank, offering a basis for SAR studies. One concern with this approach is that

lax criteria for binding and essentiality evidence may cause the list of 867 "potentially druggable" candidate targets to contain false positives. Many AlphaFill-predicted ligand hits were generic non-drug-like molecules such as ATP; more sophisticated filtering of AlphaFill hits based on chemical properties may improve the predictiveness of this strategy. In addition, few crystal structures exist for *Plasmodium* and apicomplexan parasites compared to other organisms, such as mouse or human; therefore, predicted *P. falciparum* transplant hits found with distant orthologs may not be relevant for malaria parasites, as reflected in low "druggability" scores during expert evaluation.

To minimize these issues, at the cost of deprioritizing completely novel *Plasmodium*-specific candidate targets, we focused on proteins with additional sources of binding evidence such as validated inhibitors in other species. Further validation of predicted ligand "transplants" with putative *P. falciparum* protein targets is needed. It may take several months of SAR to improve affinity strength and inhibitory potency. On the other hand, some proteins with entirely novel modes of binding may be absent from the candidate set if they lack a clear binding pocket or predicted ligand, but are in fact ligandable via a cryptic pocket, i.e., one absent in crystal structures but apparent upon binding of the right ligand. Such cryptic pockets may enable targets in protein classes historically considered undruggable, as in the case of the mutant K-Ras inhibitors[153, 154]. Molecular dynamics and/or deep learning approaches to binding pocket prediction may rescue potential false negatives[155-160].

The target evaluation rubric in this study favored proteins with prior characterization and assay development. Due to a focus on "low-hanging fruit", genes fulfilling alternative criteria, such as hitherto unexplored target classes or *Plasmodium*-specific genes of unknown function, were not highlighted by our ranking. Nevertheless, essential genes with confidently predicted

binding hit(s) provide an initial hint that may result in novel target classes, though substantial work is needed since they lack key target fulfillment data.

To date, clinically effective antimalarials with known mechanisms have been limited to drugs targeting known druggable proteins, i.e., those with well-defined, hydrophobic pockets that bind small molecule ligands. Our study therefore focused on classic druggable proteins, which are more likely to yield small molecule inhibitors with favorable drug-like characteristics. However, new approaches targeting "undruggable" proteins have emerged, such as allosteric inhibitors modulating protein-protein interactions, RNA antisense oligonucleotides (RNAi), and PROTAC (proteolysis-targeting chimera) technology[161]. Thus, it is possible that essential *P. falciparum* genes lacking small molecule-binding evidence in our analysis could be targeted through alternative methods.

We believe the list of candidates proposed in this work can serve as a starting point for future phenotypic validation and small molecule optimization efforts. As discoveries about protein structure and gene function emerge, automated data extraction and integration will be the next step toward a dynamic resource for prioritizing novel antimalarial targets. As exemplified by a similar target ranking for *Mycobacterium tuberculosis*[114], this target evaluation approach can be applied to other disease-causing organisms. For *P. falciparum* malaria, our data compendium may assist in prioritizing genes for other use cases, such as vaccine development. Furthermore, our data compendium allowed us to compare characteristics between the list of 1,929 genes with essentiality evidence and known drug targets, highlighting those with higher chances of progressing into therapeutic stages that could be obscured otherwise. Overall, we believe this project will assist the malaria community in redirecting resources and effort towards future high-quality drug targets.

3.4 Methods

3.4.1 Data acquisition

List of genes and genomic features (GFF) for *Plasmodium falciparum* 3D7 genome (PlasmoDB release 66) was downloaded and protein-coding genes were extracted along with their gene annotations and genomic location. Additional genomic annotations were obtained by querying PlasmoDB to extract UniProt and Entrez ID(s), ortholog group (OrthoMCL), protein features (CDS and protein length, molecular weight, isoelectric point), domain annotations (InterPro, Pfam, and Superfamily), number of transmembrane (TM) domains, and enzyme commission (EC) numbers. Gene function (Gene Ontology components, functions, and processes) was extracted either from PlasmoDB or by querying the InterPro ID with the InterPro2GO mapping tool from EMBL-EBI. Gene essentiality data was obtained for *P. falciparum*[124] and *P. berghei*[125, 126] parasites mapped to their *falciparum* orthologs using OrthoMCL orthology group IDs. Protein Data Bank (PDB) IDs of crystal structures were obtained by searching either gene symbols, UniProt IDs associated with each gene, or the term "*Plasmodium*". A report with gene identifier, organism, accession number, method for structure determination and publication information was extracted for the search hits.

3.4.2 Mapping genes to associated literature publications

A download from the NCBI FTP site was performed for gene2pubmed.gz (version 2024-02-21) containing taxonomy ID, gene ID (Entrez) and PubMed ID information. Gene IDs were mapped to the *P. falciparum* 3D7 annotation set, and corresponding PMIDs were extracted. To include literature references associated with gene symbols, we queried each symbol in PubMed

using the Eutils[162] efetch function from NCBI; additional information for each publication was obtained pragmatically using the same tool, namely title, authors and digital object identifier (DOI). Literature references from gene nomenclature extraction were manually reviewed and filtered for unrelated records (e.g., same name but different meaning across organisms/diseases).

3.4.3 Determining candidate proteins with evidence of small molecule binding

BindingDB[122] (version 2024-01-01) was queried to extract a list of 6,202 unique UniProt IDs with at least one ligand having a measured affinity of at least 10 μM. Ligand SMILES were extracted for target hits. These proteins were queried for sequence similarity against the OrthoMCL[127] (v.6.19) database using BLAST v2.15 blastp function[134]. Orthology of *P. falciparum* 3D7 proteins to any of the 6,202 BindingDB proteins was determined based on presence in the same ortholog group according to OrthoMCL, HOGENOM[133], OMA[129] or OrthoDB[128] phylogenomic databases, using the UniProt ID mapping tool (accessed February 2, 2024). Either direct orthology to a BindingDB protein based on at least one phylogenomic database or a BLAST hit with *E* value < 1 was considered as binding evidence based on BindingDB.

Predictions of ligands corresponding to Pf3D7 AlphaFold (v4) models were taken from the AlphaFill databank[121], which identifies candidate ligands by searching for sequence homologs in PDB[132] with known ligands and "transplanting" ligands in regions of local structural homology. AlphaFill excludes common crystallization agents such as polyethylene glycol; in order to focus on AlphaFill hits that are more likely to indicate druggability, we further excluded small ligands with less than ten atoms as well as additional salts, solvents and polymers used for protein crystallization (PDB ligand IDs: 1BO, ACN, ACT, CCN, CIT, CL, DIO, DMS, EOH, FLC, FMT, GBL, HEZ, IPA, JEF, MLA, MLI, MPD, PDO, PEG, PO4, POL, SBT, SIN, SO4, TBU, TLA)

listed in McPherson and Gavira 2014[163]. AlphaFill hits having global RMSD < 10 (structural similarity between the protein of interest and its potential homolog) and local RMSD < 4 (structural similarity of the backbone atoms within 6 Å from the transplanted ligand, after local structural alignment) were considered "confident" hits. Any Pf3D7 protein with at least one confident AlphaFill hit (global RMSD < 10 and local RMSD < 4) to a ligand satisfying the exclusion criteria was classified as having binding evidence based on AlphaFill.

Lastly, inhibitors linked to EC number classes were obtained from BRENDA Enzyme Database[123] (release 2023.1) by querying EC number annotations for Pf3D7 genes, applicable only to enzymes. Additional ligand types were not considered and for genes with incomplete EC number annotations, all EC numbers matching wildcards were considered. Each Pf3D7 gene with at least one BRENDA EC inhibitor, excluding single-atom ions, was classified as having binding evidence based on BRENDA. Classifications of binding evidence based on orthology or sequence homology to a ligandable protein in BindingDB, presence of confident AlphaFill hit(s), and presence of relevant BRENDA EC inhibitor(s) are listed for each Pf3D7 gene in **Supplemental Table 3.1**.

3.4.4 Identification of human orthologs

Homo sapiens genes (GRCh38, release 39) orthologous to Pf3D7 genes were determined based on OrthoMCL ortholog groups. Sequence and structural similarity were evaluated through pairwise comparison of Pf3D7 and human ortholog AlphaFold (v4) structures with TM-align[144].

3.4.5 Definition of hypervariable and core genomic regions

Initial definitions of hypervariable and core regions in the *P.* falciparum 3D7 genome from Miles et al. 2016[65] were adjusted on a gene-by-gene basis to include most *var*, *rifin*, *stevor*, and *Pfmc-2TM* multigene family members within subtelomeric or internal hypervariable regions. Non-nuclear genome genes were classified according to their respective chromosome (apicoplast or mitochondrial). The genome classifications for each Pf3D7 gene are listed in **Supplemental Table 3.1**.

3.4.6 Categorization of gene essentiality evidence

For each gene, evidence for blood-stage essentiality (Zhang et al., PlasmoGEM, and RMgmDB) was classified as either confidently essential, confidently nonessential, unclear, or "no data" if unavailable. Conservative thresholds were used to heuristically categorize genes as confidently essential or nonessential. In the case of the Zhang et al. *piggyBac* insertion mutagenesis dataset, which reports number of transposon insertions in addition to a Mutagenesis Index Score (MIS), genes labelled with the "Non-Mutable in CDS" phenotype were considered confidently essential if 0 insertions were observed, MIS > 0.8, and the phenotype was not noted as "tentative." Genes labelled as "Mutable in CDS" were considered confidently nonessential if number of insertions ≥ 1, MIS < 0.5, and the phenotype was not noted as "tentative". Genes measured by the Zhang et al. dataset that did not fulfill either sets of criteria were categorized as having unclear evidence of essentiality. For PlasmoGEM, genes labelled as "Insufficient data" were included in the "no data" category. A more complex classification scheme was used to rescue essential genes with a "Slow" phenotype by accounting for relative growth rate. If more than 10% or 20% of the

95% confidence interval for relative growth rate fell below 0.5 for genes labelled “Essential” or “Slow”, respectively, or the PlasmoGEM confidence score < 3 for genes labelled “Essential”, evidence was considered unclear; otherwise, genes with the “Essential” phenotype or “Slow” phenotype with relative growth rate < 0.5 were categorized as confidently essential. Meanwhile, genes labelled “Slow” with relative growth rate ≥ 0.5 were confidently nonessential if the lower bound on relative growth rate > 0.6, genes labelled “Dispensable” were confidently nonessential if either the lower bound on relative growth rate ≥ 0.5 or confidence > 3, and genes labelled “Fast” (suggesting increased growth rate upon disruption) were unilaterally considered nonessential. Finally, for RMgmDB, when the phenotype was not “nt” (not tested), evidence was categorized as confidently essential if there was a change in phenotype upon gene modification; if no difference was observed, RMgmDB evidence was considered unclear. Information from the three data sources was integrated to determine gene essentiality bins, which were “full” if all available sources suggest the gene is confidently essential, “anti” if at least one source suggests the gene is confidently nonessential, “partial” if all sources of evidence are unclear, and “no data” if the gene was not tested in any of the three datasets.

3.4.7 Assessment of gene expression by lifecycle stage

To assess gene expression in the asexual blood-stage, genes were categorized based on strength of evidence from the Le Roch et al. microarray dataset[141], which reports expression and logP values for six ABS substages synchronized using two different methods. Previous *P. falciparum* 3D7 gene IDs were mapped to current IDs using PlasmoDB. Genes were considered expressed in ABS if at least one substage showed expression value ≥ 30 and logP ≤ -1, not expressed in ABS if all substages showed expression < 10 or logP > –0.5. Otherwise, evidence

was considered unclear; such genes were labelled "potentially expressed" in ABS. The Malaria Cell Atlas Chromium 10x RNA-seq dataset was also incorporated in **Supplemental Table 3.3**; among the four stages tested (ring, trophozoite, schizont, and gametocyte), genes were considered expressed if median expression > 0, and evidence of expression was considered unclear if the third quartile of RNA-seq expression across cells > 0.

3.4.8 Candidate target scoring rubric

Ten categories belonging to each data type collected were defined to provide a total of 100 points. Availability of recombinant or in situ protein expression was scored between 0 (no information) and 4 (available). Quality of literature and quality of gene essentiality were scored between 0 and 6 points, with higher scores for greater amount of relevant literature or confident essentiality evidence. For selectivity, 0 was given if the *Plasmodium* protein and human ortholog were very similar (though exact similarity percentages varied between reviewers, a 25–83% range was observed) and data suggested selectivity could be an issue; a maximum score of 6 was given if a human ortholog was absent or there were clear differences between small molecule inhibitors for *Plasmodium* and human enzymes. Evidence of multistage expression received a score of 3 if there was evidence in one stage, up to 11 for three or more stages. Target novelty, conservation among species (genetic variation), and assay development ranged from 0 to 11 depending on prior characterization of the protein, especially as a potential antimalarial target (novelty); extent of conservation (for genetic variation); or availability of binding/functional assays. Structural information scored from 0 to 17 depending on whether crystal structures were available for orthologs in other organisms, *Plasmodium* proteins, or as ligand-bound structures. Lastly,

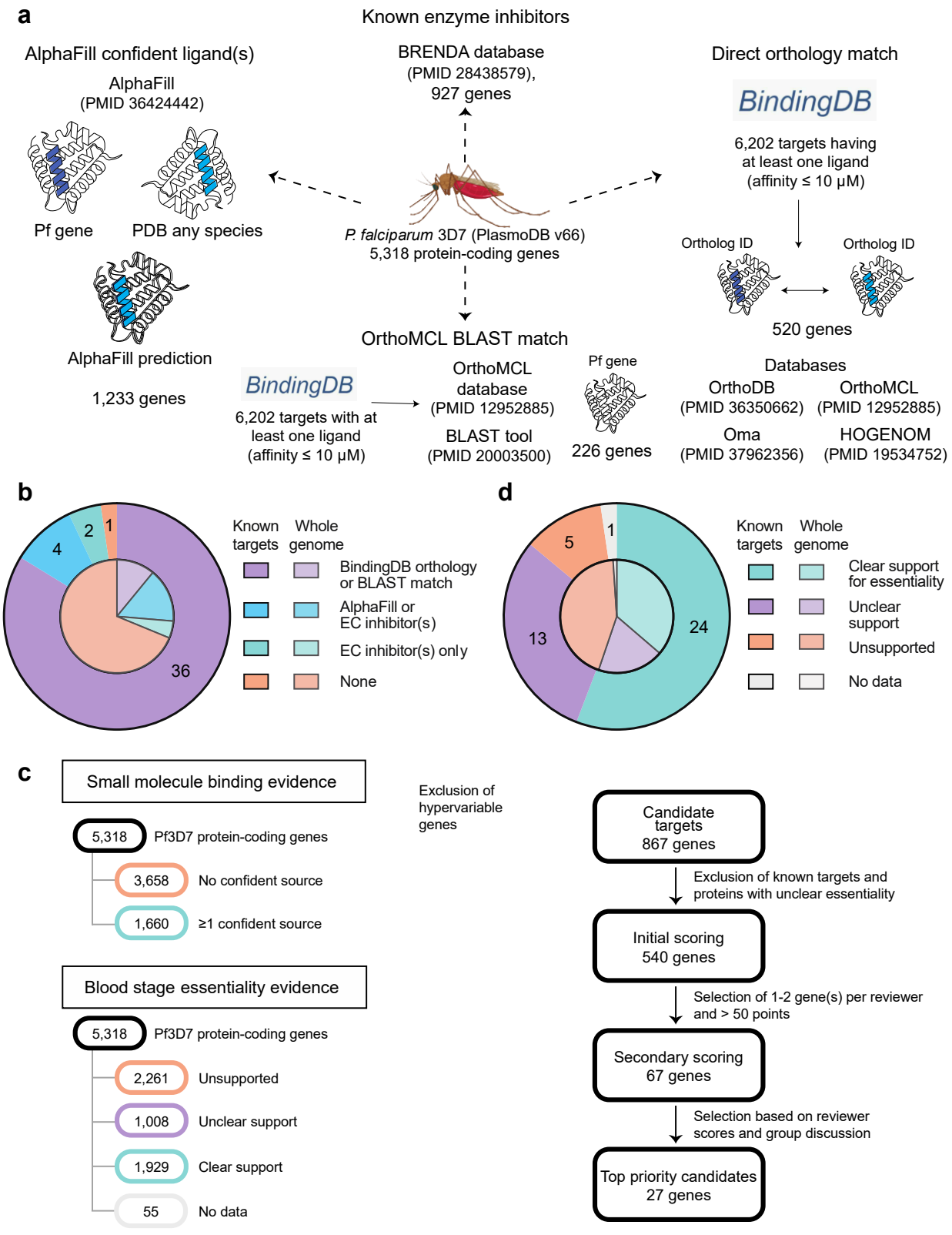
druggability was scored from 0 (no evidence of binding pocket) to 17 points if a tool compound inhibiting *Plasmodium* growth was known.

3.5 Acknowledgements

Chapter 3 is a partial, modified reprint of the material as it appears in NPJ Drug Discovery, 2025. Karla P. Godinez-Macias, Daisy Chen, J. Lincoln Wallis, Miles G. Siegel, Anna Adam, Selina Bopp, Krypton Carolino, Lauren B. Coulson, Greg Durst, Vandana Thathy, Lisl Eshrick, Madeline A. Farringer, Erika L. Flannery, Barbara Forte, Tiqing Liu, Luma Godoy Magalhaes, Anil K. Gupta, Eva S. Istvan, Tiantian Jiang, Krittikorn Kumpornsin, Karen Lobb, Kyle J. McLean, Igor M.R. Moura, John Okombo, Connor N. Payne, Andrew Plater, Srinivasa P.S. Rao, Jair L. De Siqueira-Neto, Bente A. Somsen, Robert L.Summers, Rumin Zhang, Michael K. Gilson, Francisco-Javier Gamo, Brice Campo, Beatriz Baragaña, James Duffy, Ian H. Gilbert, Amanda K. Lukens, Koen J. Dechering, Jacquin C. Niles, Case W. McNamara, Xiu Cheng, Lyn-Marie Birkholtz, Alfred W. Bronkhorst, David A. Fidock, Dyann F. Wirth, Daniel E. Goldberg, Marcus C.S. Lee, Elizabeth A. Winzeler. "Revisiting the *Plasmodium falciparum* druggable genome using predicted structures and data mining". The dissertation author was a co-first author of this paper.

Figure 3.1 Identification of *P. falciparum* protein-coding genes that are potentially druggable (*n* = 1,660) and genes that have evidence of blood-stage essentiality (*n* = 2,992)

(A) Identification of potentially druggable genes. Diagram illustrating the four methods used to identify genes with evidence of small molecule binding. Out of the 5,318 protein-coding 3D7 *falciparum* genes, 226 were homologous to at least one of the 6,202 validated drug targets in BindingDB based on BLAST, while 520 genes were found to be orthologous based on the phylogenomic databases OrthoMCL, OrthoDB, OMA or HOGENOM. We found 1,233 genes with confident AlphaFill hit transplant(s), and 927 *falciparum* genes were mapped to EC numbers with inhibitors in the BRENDA database. Mosquito graphic was generated with BioRender (https://BioRender.com/s17b944). **(B)** Binding evidence for validated targets vs. all genes. Distribution of binding evidence for 43 known targets[67] (outer pie) compared to the distribution across all 5,275 *P. falciparum* 3D7 genes (inner pie). **(C)** Workflow for the identification of 867 candidate targets and subsequent prioritization. Identification and filtering process to define *P. falciparum* candidate drug targets. The intersection of 1,660 genes with evidence of small molecule binding and 2,992 genes with essentiality support yielded 867 candidates, after excluding hypervariable regions. Subsequent candidate prioritization resulted in 540 candidates subjected to initial scoring, 67 of which underwent a second round of scoring, culminating in 27 top ranking targets that were discussed by a panel of experts. **(D)** Essentiality classifications for validated targets vs. all genes. Distribution of essentiality classifications for 43 known targets[67] (outer pie) compared to the distribution for 5,275 3D7 genes (inner pie). Essentiality classifications are described in **methods**.


a
Known enzyme inhibitors
AlphaFill confident ligand(s)
AlphaFill
(PMID 36424442)
Pf gene
PDB any species
AlphaFill prediction
1,233 genes
BRENDA database
(PMID 28438579),
927 genes
P. falciparum 3D7 (PlasmoDB v66)
5,318 protein-coding genes
OrthoMCL BLAST match
BindingDB
6,202 targets with at least one ligand (affinity ≤ 10 µM)
OrthoMCL database
(PMID 12952885)
BLAST tool
(PMID 20003500)
Pf gene
226 genes
Direct orthology match
BindingDB
6,202 targets having at least one ligand (affinity ≤ 10 µM)
Ortholog ID
Ortholog ID
520 genes
Databases
OrthoDB
(PMID 36350662)
OrthoMCL
(PMID 12952885)
Oma
(PMID 37962356)
HOGENOM
(PMID 19534752)
b
Known targets
Whole genome
BindingDB orthology or BLAST match
AlphaFill or EC inhibitor(s)
EC inhibitor(s) only
None
4
2
1
36
d
Known targets
Whole genome
Clear support for essentiality
Unclear support
Unsupported
No data
5
1
13
24
c
Small molecule binding evidence
5,318 Pf3D7 protein-coding genes
3,658 No confident source
1,660 ≥1 confident source
Blood stage essentiality evidence
5,318 Pf3D7 protein-coding genes
2,261 Unsupported
1,008 Unclear support
1,929 Clear support
55 No data
Exclusion of hypervariable genes
Candidate targets
867 genes
Exclusion of known targets and proteins with unclear essentiality
Initial scoring
540 genes
Selection of 1-2 gene(s) per reviewer and > 50 points
Secondary scoring
67 genes
Selection based on reviewer scores and group discussion
Top priority candidates
27 genes

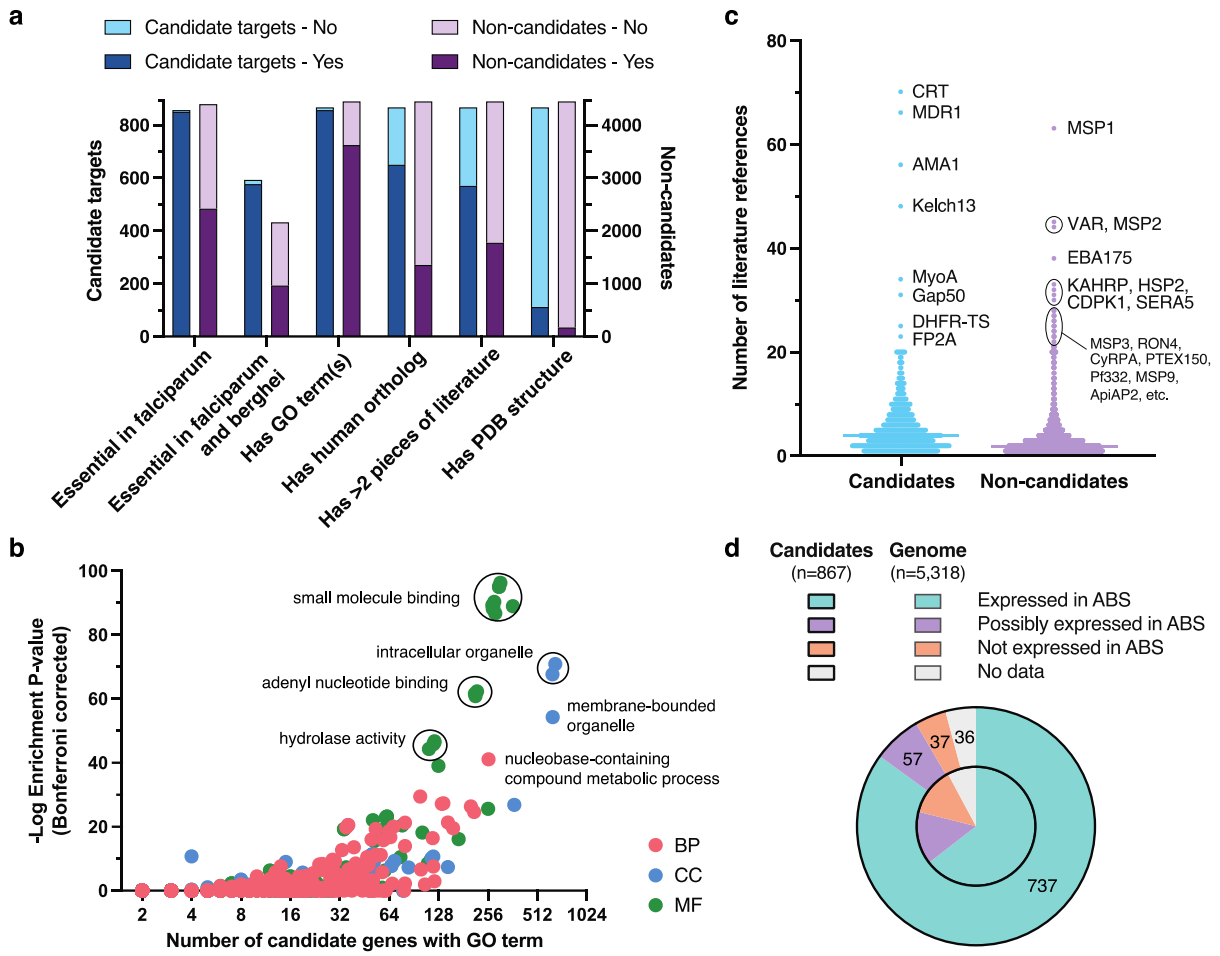


Figure 3.2 Characteristics of 867 candidate targets compared to 4,451 non-candidate genes

(A) Comparison of characteristics between the 867 candidates (blue) and 4,451 non-candidates (purple). Numbers of candidate targets versus non-candidates labelled essential only in the *P. falciparum* essentiality screen and not the *P. berghei* datasets (850 vs. 2,421); labelled essential in both *P. falciparum* and *berghei* datasets (577 vs. 967); having at least one GO term (857 vs. 3,624); having human ortholog(s) (650 vs. 1,356); having > 2 associated literature references (570 vs. 1,775); or having PDB structures (112 vs. 174) are shown. **(B)** GO term enrichment analysis for the 867 candidate targets compared to all 5,318 protein-coding genes in the *P. falciparum* 3D7 genome using GOATOOLS[164]. Terms with ontology tree depth > 2 (n = 939) are displayed based on number of candidates having the GO term (X axis) versus $-\log_{10}$ Bonferroni corrected enrichment P value, with a maximum uncorrected P value of 0.05 (Y axis, Fisher's exact test). Points are colored by ontology type: red for biological process (BP), blue for cellular component (CC), and green for molecular function (MF). Highly enriched terms are labeled. **(C)** Distributions of number of unique scientific publications associated with the 867 candidate targets (blue) vs. 4,451 non-candidate genes (purple). Median lines are shown, and the most highly referenced genes are labeled. **(D)** Distribution of classifications for gene expression evidence in ABS according to Le Roch et al.[141]. Candidate targets with clear evidence of ABS expression (737, teal), unclear evidence (57, purple), no expression (37, orange) and no data (36, grey) are shown in the outer pie, in contrast to the distribution among all *P. falciparum* protein-coding genes in the inner pie.

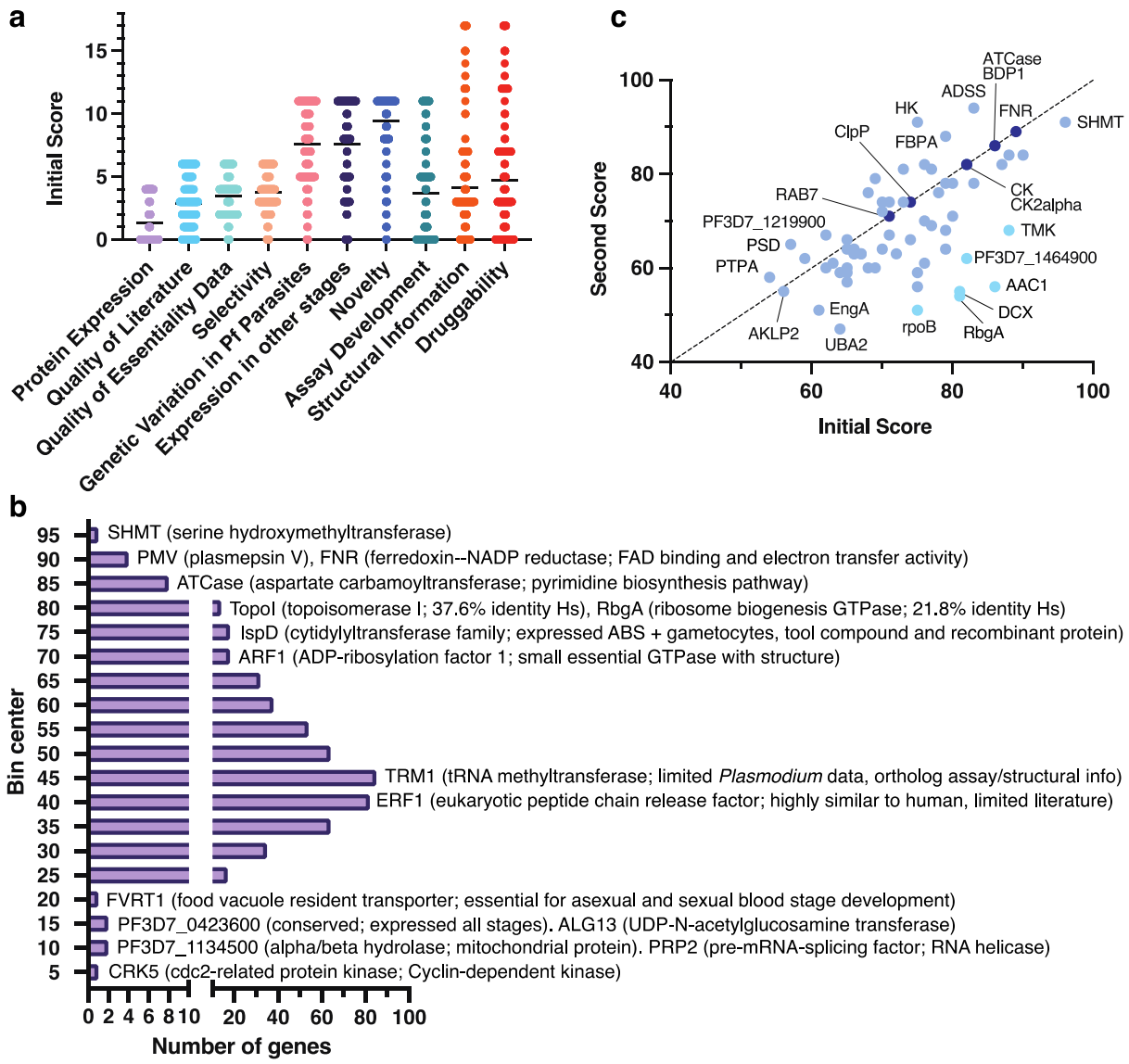


Figure 3.3 Rubric-based scoring of 540 candidate targets with strong evidence of essentiality

(A) Gene scores across rubric categories. Plot showing first score distributions for 540 candidate targets with strong evidence of blood-stage essentiality across the ten categories of the scoring rubric (**methods**). Average value per category is shown for each category. **(B)** Frequency distribution of first total scores. Histogram showing the frequency (*X* axis) of first total scores (*Y* axis) for the 540 scored candidates. Bin center was determined and plotted with Prism v.9.5. Examples of candidates falling in select total score bins are shown, including gene product description and notable characteristics when applicable. **(C)** Comparison of first and second scores for the top 67 scored candidates subject to secondary review. Identity line is marked by a dashed black line. Dark blue circles denote equal score in both rounds; light blue circles represent score differences of 1–19 points; and cyan circles represent a score difference of at least 20 points.

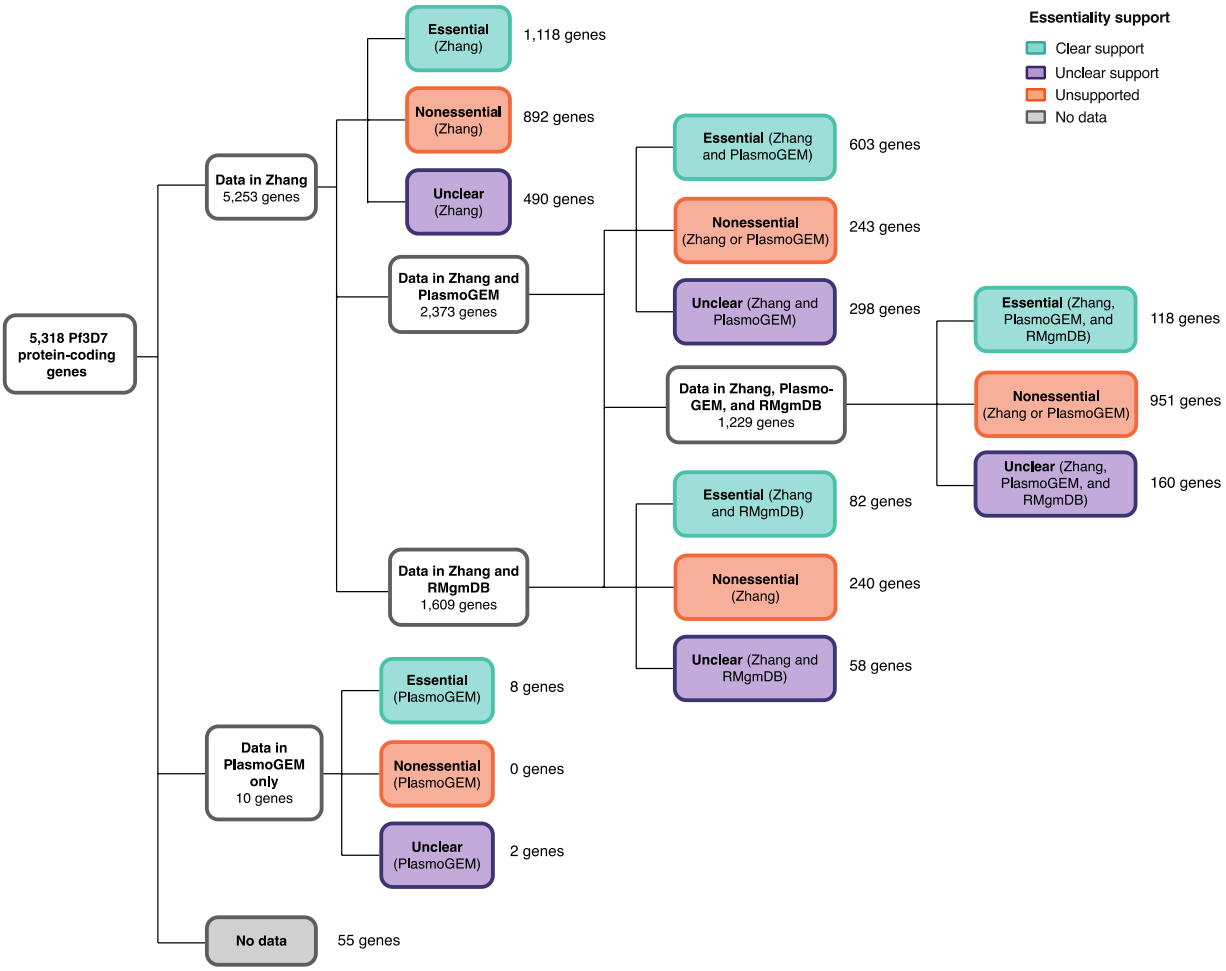
Table 3.1 List of 67 "druggable target" candidates from the entire *P. falciparum* genome.

Ranking scores for 67 attractive druggable candidate targets. Total scores determined by reviewers for first and second round are provided (see **methods**). Highly attractive gene targets are indicated under "selected" column.

Gene ID	Gene name	Gene description	Score 1	Score 2	Selected
PF3D7_0206700	ADSL	adenylosuccinate lyase	80	78	yes
PF3D7_0209300	IspF	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	76	82	yes
PF3D7_0307400	ClpP	ATP-dependent Clp protease proteolytic subunit	74	74	yes
PF3D7_0410700	RbgA	ribosome biogenesis GTPase A, putative	81	54	yes
PF3D7_0510500	TopoI	topoisomerase I	80	71	yes
PF3D7_0623200	FNR	ferredoxin--NADP reductase	89	89	yes
PF3D7_0624000	HK	hexokinase	75	91	yes
PF3D7_0629000	GNA1	glucosamine 6-phosphate N-acetyltransferase	65	64	yes
PF3D7_0907900	PDF	peptide deformylase	83	78	yes
PF3D7_1012400	HGPRT	hypoxanthine-guanine phosphoribosyltransferase	73	81	yes
PF3D7_1033700	BDP1	bromodomain protein 1	86	86	yes
PF3D7_1108400	CK2α	casein kinase 2, alpha subunit	82	82	yes
PF3D7_1113100	PRL	protein tyrosine phosphatase	70	74	yes
PF3D7_1120100	PGM1	phosphoglycerate mutase, putative	79	68	yes
PF3D7_1144900	RAB6	ras-related protein Rab-6	76	61	yes
PF3D7_1235600	SHMT	serine hydroxymethyltransferase	96	91	yes
PF3D7_1239500	GyrB	DNA gyrase subunit B	76	70	yes
PF3D7_1247400	FKBP35	peptidyl-prolyl cis-trans isomerase FKBP35	71	74	yes
PF3D7_1251300	TMK	thymidylate kinase	88	68	yes
PF3D7_1320100	ClpS	ATP-dependent Clp protease adapter protein ClpS	77	69	yes
PF3D7_1344800	ATCase	aspartate carbamoyltransferase	86	86	yes
PF3D7_1354500	ADSS	adenylosuccinate synthetase	83	94	yes
PF3D7_1444800	FBPA	fructose-bisphosphate aldolase	79	88	yes
PF3D7_1446200	LAP	M17 leucyl aminopeptidase	87	82	yes
PF3D7_1453800	GluPho	glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase	79	78	yes
PF3D7_1460400	UCHL3	ubiquitin carboxyl-terminal hydrolase isozyme L3	88	84	yes
PF3D7_0106900	IspD	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, putative	74	66	
PF3D7_0202700	OPP	octaprenyl pyrophosphate synthase	59	62	
PF3D7_0512700	OPRT	orotate phosphoribosyltransferase	63	61	
PF3D7_0514300		aspartate--tRNA ligase, putative	65	60	
PF3D7_0517800	DCX	apicortin	81	55	
PF3D7_0616000	PDXK	pyridoxal kinase	68	76	
PF3D7_0708800	HSP110c	heat shock protein 110	62	60	

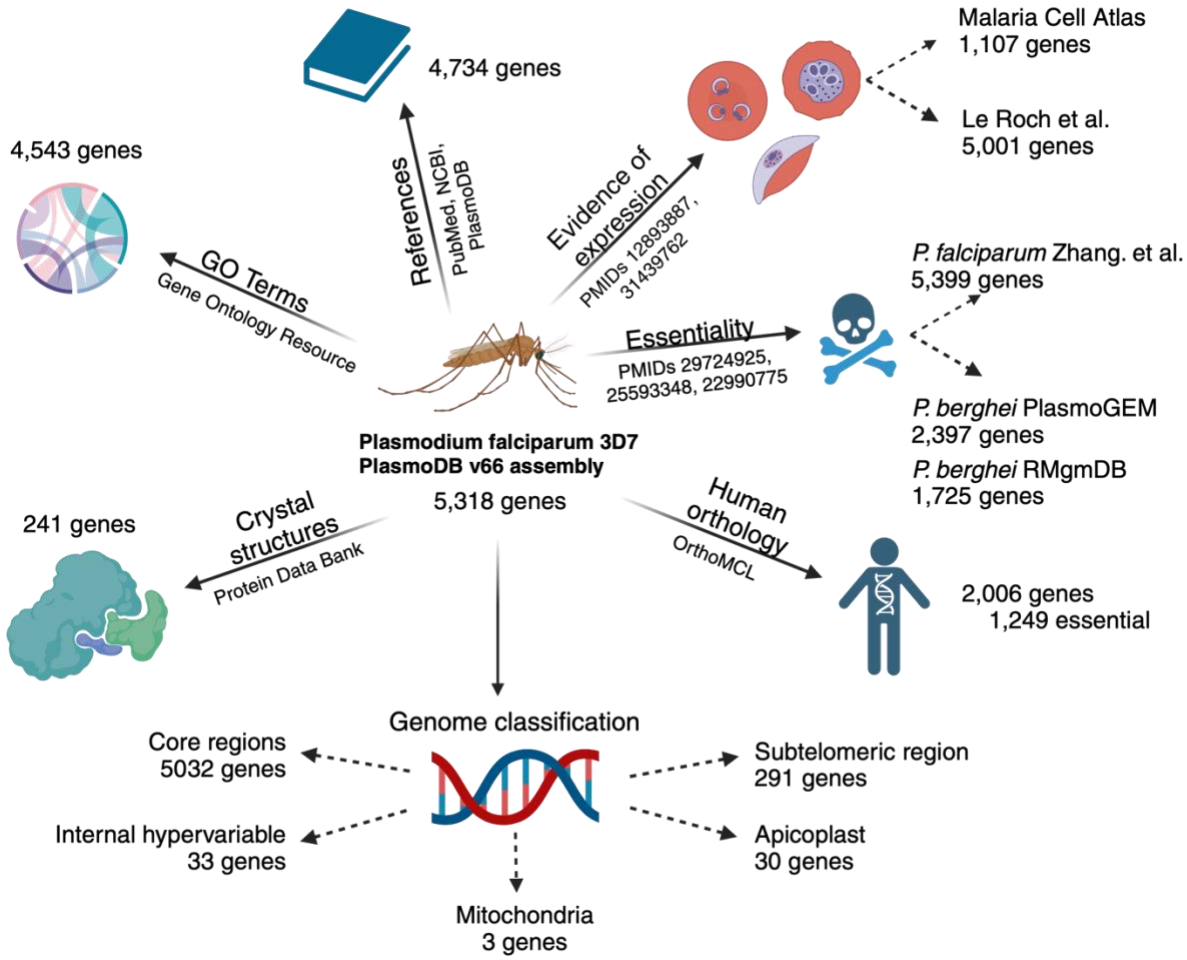
Table 3.1 List of 67 "druggable target" candidates from the entire *P. falciparum* genome, Continued.

Gene ID	Gene name	Gene description	Score 1	Score 2	Selected
PF3D7_0903200	RAB7	ras-related protein RAB7	71	71	
PF3D7_0927900	PSD	phosphatidylserine decarboxylase	57	65	
PF3D7_0931900	AKLP2	adenylate kinase-like protein 2	56	55	
PF3D7_1004000		60S ribosomal protein L13, putative	75	59	
PF3D7_1008700		tubulin beta chain	77	81	
PF3D7_1014700	PHB2	prohibitin 2	66	63	
PF3D7_1020900	ARF1	ADP-ribosylation factor 1	70	72	
PF3D7_1022800	ISPG	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase (ferredoxin)	65	57	
PF3D7_1037300	AAC1	ADP, ATP carrier protein 1	86	56	
PF3D7_1118600	MYST	histone acetyltransferase MYST	69	79	
PF3D7_1136500	CK1	casein kinase 1	72	63	
PF3D7_1202000	PSH2	ATP-dependent DNA/RNA helicase PSH2	69	60	
PF3D7_1217300	EngA	GTP-binding protein EngA	61	51	
PF3D7_1219900		ribulose-phosphate 3-epimerase, putative	62	67	
PF3D7_1223100	PKAr	cAMP-dependent protein kinase regulatory subunit	78	76	
PF3D7_1223300	GyrA	DNA gyrase subunit A	73	74	
PF3D7_1230400	ClpQ	ATP-dependent protease subunit ClpQ	64	59	
PF3D7_1237000	UBA2	SUMO-activating enzyme subunit 2	64	47	
PF3D7_1323500	PMV	plasmepsin V	90	84	
PF3D7_1324900	LDH	L-lactate dehydrogenase	71	67	
PF3D7_1330600		elongation factor Tu, putative	79	64	
PF3D7_1333200	UBA1	ubiquitin-activating enzyme	65	66	
PF3D7_1337800	CDPK5	calcium-dependent protein kinase 5	67	63	
PF3D7_1340600	DBR1	RNA lariat debranching enzyme, putative	75	56	
PF3D7_1401800	CK	choline kinase	82	82	
PF3D7_1414600	Pgt1	mRNA-capping enzyme subunit alpha	70	64	
PF3D7_1415000	UDG	uracil-DNA glycosylase	67	63	
PF3D7_1415800	KsgA1	ribosomal RNA small subunit methyltransferase A1	68	60	
PF3D7_1417800	MCM2	DNA replication licensing factor MCM2	66	63	
PF3D7_1430100	PTPA	serine/threonine protein phosphatase 2A activator	54	58	
PF3D7_1438500	CPSF3	cleavage and polyadenylation specificity factor subunit 3, putative	66	64	
PF3D7_1440300	PBGS	porphobilinogen synthase	65	59	
PF3D7_1464900		ATP-dependent zinc metalloprotease FTSH, putative	82	62	
PF3D7_API04400	rpoB	DNA-directed RNA polymerase subunit beta, putative	75	51	



Supplemental Figure 3.1 Essentiality categorization determination scheme for 5,318 *P. falciparum* protein-coding genes for target identification

Criteria to categorize genes in essentiality bins, according to supporting data for essentiality in the asexual blood-stage from different datasets[124-126]. Four categories (clear support, unclear support, unsupported, and no data) were defined. Genes under each category based on available data are shown with teal squares for the 1,929 genes having clear essentiality support, purple squares for the 1,008 genes having unclear unsupport, and orange squares for the 2,326 genes categorized as clearly unsupported (e.g., non-essential genes). The remaining 55 genes, in grey, are those lacking essentiality data. Breakdown of genes with support across the datasets is shown.



Supplemental Figure 3.2 Data collected for full genome

Information collected for all 5,318 *P. falciparum* genes from PlasmoDB release 66[131]. Database resources for collected information are shown[65, 127, 130, 131, 139, 140, 144]. Total number of genes having at least one record per category is indicated. Out of the 19,900 genes in the human genome, there were 2,006 orthologous *P. falciparum* genes, of which 1,249 were determined to be essential. Genome classification for the 5,318 genes was determined according to **methods**. Figure was created in BioRender (Godinez, K. (2025) https://BioRender.com/q39w386).

Chapter 4

Dissecting imidazolopiperazine and naphthoquinone mechanisms of action in *Plasmodium* asexual malaria parasites

4.1 Introduction

Phenotypic screening is a powerful method to quantify the inhibitory response in a cell or organism[165]. In malaria parasites, many potent small molecule inhibitors have been identified through this method, with some advancing into late stages of clinical trials (e.g., cipargamin and ganaplacide)[66, 166, 167]. A limitation, however, is the lack of information regarding the mechanism of action (MoA)[66]. While not imperative, this knowledge enables a more efficacious drug design and use[66, 168]. For this reason, only the most potent hits from a phenotypic screen are further characterized with in vitro and whole genome sequencing, proteomics, biophysical or other target identification assays[166]. Nevertheless, these studies may result in identification of off-target effects instead of the actual drug-target[166], remaining with the drug-target pair unresolved. Thus, as new uncharacterized antimalarial compounds become available, there is a need for new target identification methods.

Gene expression profiling has mainly been used to elucidate transcriptional regulatory networks and to functionally assign uncharacterized genes to networks in many species[169-171]. In theory, transcriptional profiling could be used to determine a target or mechanism of action of an uncharacterized small molecule[172, 173] especially in cases where a metabolic block leads the parasite to compensate by transcriptionally upregulating the genes in the blocked pathway[174] (e.g., upregulation of the folate biosynthetic pathway after treatment with an antifolate). This principle has been demonstrated by Tewari et al.[175], where they show that *Plasmodium falciparum* parasites, causatives of malaria, when treated with cipargamin (a spiroindolone) and PA21A092 (a pyrazoleamide) differ in transcriptional profile and response but both show upregulation of alpha–beta hydrolase, an enzyme that hydrolyzes ATP; this finding is consistent with cipargamin and PA21A092 targeting *Pf*ATP4 (P-type ATPase)[175]. In addition, even if a specific pathway is not

revealed, transcriptional profiles could be used to categorize drugs with the same MoA[173] (e.g., all antifolates should give similar antifolate transcriptional profiles). Challenges with applying transcriptional profiling to malaria parasites include problems with synchronization of cells and determining the optimum cell-type to conduct expression profiling[176]. Furthermore, determination of time and concentration of drug exposure, in the case of treatment response, adds an extra layer of liability. In theory, with single-cell we could overcome this problem by allowing gene expression detection in a variety of different cell-types all at once[176].

To assess the use of single-cell sequencing as a method to provide clues about the target or mechanisms of action for a small molecule, we performed profiling on human red blood cells infected with *P. falciparum* Dd2 parasites, that were subsequently treated with one of the known antimalarials artemisinin, atovaquone, and gnf179. Atovaquone is a naphthoquinone molecule targeting *Plasmodium* cytochrome bc-1 complex[66] as part of the respiratory chain complex; recently, it was shown that the imidazolopiperazine (gnf179) scaffold mechanisms is related to protein trafficking and secretion pathways[177], with *pfsey1* (synthetic enhancer of YOP1) likely playing a significant role[178]; lastly, while no definitive target is known for artemisinin (sesquiterpene lactone class)[66], this compound activates free radicals in the presence of heme, leading to the parasite's death[179]. Through a custom bioinformatics pipeline, we distinguished cell-types across the parasite's lifecycle and observed different transcriptional patterns relative to the antimalarial when comparing drug-treated versus non-treated parasites. To identify the molecular differences between treated and non-treated parasites, we performed differential gene expression analysis, highlighting 2,239 genes upregulated when treated with atovaquone, 3,172 genes with artemisinin, and 1,488 genes with gnf179. A pathway analysis of these genes showed

enrichment in the electron transport chain and sphingolipid metabolism pathways for atovaquone and gnf179, respectively. This finding is consistent with the literature, suggesting an optimistic use of this method to investigate mechanisms of action for uncharacterized small molecules.

4.2 Results

4.2.1 Examining drug treatment during *Plasmodium* falciparum asexual blood-stage malaria infection

To perform cell profiling, blood-stage *Plasmodium falciparum* Dd2 parasites were added to four different flasks; one flask was maintained as control and the remaining three were treated with one of three antimalarials as described in **Figure 4.1A**. Shortly, 50nM of atovaquone was added to one flask, 100nM of atovaquone to a second flask and 100nM of gnf179 was added to a third flask. After 48 hours of drug exposure, cells were extracted using magnetic separation (expect for the first replicate), obtaining an average of 1,200 infected red blood cells per sample. Samples were then prepared for barcoding, RNA library preparation and sequencing using 10x genomics technology (see **methods** for details). After sequencing, we mapped the reads to a custom reference genome of human (GRch38.p13) and *P. falciparum* Dd2 (v62) genomes. This mapping resulted in an average of 124,903.67 *Plasmodium* cells for control (non-treated) samples, 29,174 cells for artemisinin-treated samples, 91,108.33 cells for atovaquone-treated, and 93,705.00 cells for gnf179-treated samples (**Figure 4.1B**). Due to the large differences in initial cell population across samples, we randomly selected a total of 15,000 cells per sample, to balance out the imbalanced dataset and ensure similar size across comparisons[180] (**Figure 4.1B**).

To investigate transcriptional changes relative to after drug treatment, we integrated cells from treated and non-treated parasites, focusing on protein-coding genes (**Figure 4.1C**). We first integrated cells from the three replicates and performed a standard single-cell analysis workflow using Seurat[181] with batch correction using replicate number as the variation variable. Next, we integrated cells from the control and one of the drug-treated sample (artemisinin, atovaquone, and gnf179), performed a standard single-cell workflow, and performed batch correction using the condition (control, treatment) as artifact to remove sample-specific effects and ensure true biological variability. In general, we observed high-quality cells across our *Plasmodium* population (**Figure 4.1D**) and cells were integrated as expected across replicates.

4.2.2 *Plasmodium* asexual lifecycle patterns are different relative to after drug treatment

We next sought to evaluate the parasite lifecycle transcriptional changes relative to after drug treatment. To do this, we contrasted the parasite heterogeneity present across integrated samples. We obtained a total of 13 clusters for parasites treated with artemisinin with an average of 2,176 cells per cluster, atovaquone-treated parasites were distributed across ten populations with an average of 2,269 per cluster, and eleven clusters were found after gnf179 treatment having an average of 2,084 cells per cluster. To better discern which cluster are informative to the treatment response, we plotted the distribution of cells present across all clusters (**Supplemental Figure 4.1**). Not surprisingly, we found few (n = 5) overlapping clusters among all samples. A possible explanation is that parasites were sequenced while transitioning between two lifecycle stages, having similar expression profiles while undergoing developmental processes. Another explanation could be due to a technical factor during the data normalization or integration steps[182]. Since overlapping clusters showed a significantly lower proportion (n = 2.23%) of cells compared

to the rest (n = 97.77%), we could not provide enough statistical confidence and thus removed those with less than 700 cells (highest cluster n = 690) as an attempt to reduce data noise. With the remaining clusters, we identified the parasite's heterogeneity, by identifying cell-type stages corresponding to the *Plasmodium* asexual cycle, using established gene markers or gene expression relative to the stage (**Table 4.1**, and **Figure 4.2A**).

Among the three comparisons, we observed clear patterns for developmental stages of the lifecycle (i.e., rings, trophozoites, schizonts, and gametocytes) (**Figure 2A**). The only exception was with atovaquone where a clear ring cluster was not observed. Noteworthy, enrichment of sexual stages for cells treated with artemisinin was observed, which is consistent with artemisinin being less active against late stages of the asexual development[183]. In all treatments, we were not able to fully define a cluster's cell-type since a subpopulation of cells showed similar expression for markers in two development stages (e.g., trophozoites and schizont). We also observed cases where distinct clusters had the same cell-type, such as schizonts in artemisinin- and atovaquone-treated cells. Unexpectedly, all drug-treated samples had at most two clusters with neglectable differences in gene expression between markers on that cluster compared to the rest (see **methods**). Thus, we were not able to identify them and were marked as "not applicable (NA)". Although there is a lack of completeness in cell identification, the altered composition may be associated with the parasite response during drug exposure and can be further evaluated to provide clues on the drugs mechanisms.

To investigate the differences between subpopulations for the same cell-type, we compared the enrichment of gene markers present within clusters in the cell-type (**Table 4.2**). In general, we observed enrichment for parasites on stages near the one of interest (**Figure 4.2A**). For example, schizonts in cluster one during atovaquone treatment showed higher enrichment for merozoite

parasites (*pfama1*, *pfmtrap*) while cluster two showed higher expression in ring parasites (*pfetramp2*). Similarly, parasites treated with artemisinin showed that trophozoites in cluster one had enrichment in early gametocytes (*pfg27/25*, *pfd80*) and enrichment for cluster two is higher to ring parasites (*pfhgprt*). Altogether, these differences could help elucidate genes involved in parasites developmental changes and those committed to sexual development.

4.2.3 Differential expression shows distinct drug response profiles dependent on the chemotype

We next performed differential gene expression analysis to assess drug response profiles between treatments and to get clues on the mechanism of action (**Supplemental Table 4.1**). As expected, transcriptomic profiles between drug treatments were different; we obtained a total of 3,172 genes that were statistically different (up- or down-regulated, p-value ≤ 0.05) between artemisinin-treated parasites compared to control parasites, a total of 1,488 genes after gnf179 treatment and 2,239 genes were differentially expressed after atovaquone exposure. From these, there were cases where a gene was differentially expressed in two or more clusters. For example, PfDd2_130075100, a conserved protein, was upregulated (ranging from 2.37– to 7.51–fold change) during artemisinin treatment in all schizont and nearby clusters; this is consistent with gene expression changes throughout the parasite lifecycle[184].

To better visualize the most significant differences between treatments, we plotted gene expression profiles contrasting treated parasites. Consistent with the lifecycle pattern during artemisinin treatment, we observed gametocyte-genes (e.g., *pfg27-25*) to be significantly upregulated compared to the rest, confirming that parasites late in development are able to progress their cycle while in presence of the drug. We also observed, with various abundance levels,

upregulation of heat shock protein 70 that although it is one of the most highly expressed gene[185], it is also thought to be involved in stress response mechanisms[185] suggesting we could further evaluate the response profile of altered genes uniquely associated with each drug to get insight into their mechanism. In all cases we observed significant downregulation of ETRAMP proteins (ranging from -0.60– to -6.43–fold) confirming the expected efficacy of treatment in reducing the number of malaria parasites; these proteins are found in the parasitophorous vacuole membrane of the parasite thus play a significant role in development and host interactions[186].

Reasoning that the mechanism of action for each drug should result in distinct transcriptional responses, we compared genes with at least ± 3× fold-change (i.e., $\log_2$ fold change of at least 3 or -3 per treatment) to capture those with the most biological relevance (**Figure 4.2B**). Considering cases where a gene was statistically regulated in several clusters (e.g., nearby schizont clusters), we calculated the average expression level for same genes. This resulted in 281 unique genes from artemisinin treatment, 132 unique genes from gnf179 and 159 unique genes with atovaquone treatment. In general, we observed that gene expression patterns were different among the treatments. Interestingly, we notice two cases that had opposite response between chemotypes. These are the chaperone protein *Pf*CLPM (PfDd2_000009000) that was upregulated in artemisinin but downregulated in the other two, and PfDd2_040011200, a conserved protein, that was downregulated with artemisinin, upregulated in atovaquone and not differentially expressed during gnf179 exposure. This further highlights our ability to utilize transcriptional profiling after drug exposure to determine its pharmacological action.

To get a better sense of which genes are unique to the treatment, we found the intersection of transcriptional response between groups based on gene expression (**Figure 4.2C**). A total of 475 unique genes were found, having a higher proportion transcriptionally altered uniquely to the

treatment (artemisinin 78.64%, gnf179 54.54%, atovaquone 64.15%). Minority of genes (n = 80) were regulated by at least two chemotypes, with 17 shared across all treatments (**Figure 4.2C**). Among these 80, we found conserved proteins (PfDd2_010012000, 040027900, 070028900 and 130075100) and *Pf*IMC1c, an inner membrane complex protein involved in cell invasion and shape integrity[187]. Noteworthy, out of the 475 unique genes with at least ± 3× fold-change, we found six cases (PfDd2_010006300, 040012700, 110036700, 000009000, 040011200 and 000008000) where the gene's expression was conflicted within the same treatment. For example, PfDd2_110036700 (*pfama-1*) had 3.57–fold in trophozoites treated with artemisinin, while the same gene was downregulated to -3.31–fold in schizont parasites. This result highlights the importance of investigating each parasite individually, to enable a more precise landscape of genes playing a role throughout its lifecycle and how they are altered during treatment. Overall, we found significant differences in the transcriptional response relative to after treatment, suggesting our enablement to pinpoint the compound's mechanism of action.

4.2.4 Finding *Plasmodium* falciparum orthologs between Dd2 and 3D7 strains for differentially expressed genes

Reasoning that by selecting gene expression changes of 1–fold (upregulated) or -1–fold (downregulated) relative to control would provide a more comprehensive overview of the parasite's response for each drug, we focused on statistically significant genes from each treatment as identified in **Supplemental Table 4.1**. In total, we extracted 1,486 unique genes for artemisinin, 969 genes for atovaquone and 353 gnf179 upregulated genes, and 249, 197, and 893 unique downregulated genes, respectively (**Figure 4.3A**). Interestingly, we observed a higher imbalance in the number of genes positively regulated in artemisinin (n = 916), compared to the atovaquone

(n = 643) and gnf179 (n = 342); this suggests a more significant dysregulation of the parasite biology after artemisinin exposure. Overall, we obtained more upregulated genes (n = 2,808) than downregulated (n = 1,339) genes, of which 1,533 were unique upregulated and 487 were unique downregulated genes.

As a next step, for each set of transcriptionally regulated genes we identified their orthologs in *P. falciparum* 3D7, a well-studied strain from the same species. This step was performed since available pathway enrichment tools[188] allow for queries with the 3D7 strain but not Dd2 strain. Thus, to assign orthologs we took the list of 4,147 unique differentially expressed Dd2 identifiers and mapped their 3D7 ortholog based on the Orthologs and Paralogs dataset within VEuPathDB[131] (**Supplemental Table 4.1**). From these, we observed a total of 2,020 unique Dd2 genes, of which 1,506 had a 3D7 ortholog for the upregulated set among all treatments, 470 orthologs for the downregulated set, and did not find orthologs for only 44 genes from both sets (**Figure 4.3B**).

Although function similarity in some cases does not correlation between orthologs[189], we obtained an adequate correspondence between ortholog function in our matches. For example, 1,372 of the 1,976 orthologs matched had the same function between strains, and only 604 differ either because it was putatively known in one strain, exact protein characterization is available in one strain (e.g., PfDd2_030022100 is a 6-cystein protein with its ortholog, PF3D7_0317100, known as B9 from the same family), or they are distinct enzymes involved in the same processes (e.g., ATP-dependent RNA helicase DBP7 enzyme in PfDd2_070024800 and DDX31 in PF3D7_0721300). Furthermore, we observed 203 cases where an ortholog from the 3D7 strain provided some putative clues about the function of the Dd2 strain, as it is unknown in the Dd2. Some examples of this included ATPase (PF3D7_0626700) that was upregulated after treatment with atovaquone, and the female development protein FD2 (PF3D7_1146800) showing

upregulation after artemisinin exposure. Based on these results, 3D7 orthologs were thus confidently used for the pathway analysis.

4.2.5 Pathway analysis of differentially expressed genes show hints about the regulatory mechanisms of atovaquone and gnf179

To uncover the transcriptional regulatory changes occurring in *Plasmodium* parasites after drug treatment, we performed a pathway enrichment analysis (**Supplemental Table 4.2**). Taking the list of 643, 342 and 916 genes that were upregulated in atovaquone, gnf179 and artemisinin (**Supplemental Table 4.1** and **Figure 4.3B**), we used Metascape[188] with default settings to identify the pathways for which the set of genes appear at rates greater than chance for each drug, separately. This step was also repeated for the 186, 282 and 234 downregulated genes. Since the human proteins are of primary focus in Metascape, we observed that majority of the enriched pathways were a result of a human-ortholog inference that may not relate to *Plasmodium* parasites, and thus were not considered for their analysis. After filtering human-inferred pathways, we obtain enrichment (Log *p*-value < -1) for six upregulated pathways in artemisinin and gn179 and five pathways in atovaquone (**Figure 4.3C**). As expected, the lower proportion of transcriptionally downregulated genes in all treatments also resulted in just a few enriched pathways, obtaining four for artemisinin, and two pathways for gnf179 and atovaquone.

Based on previous results, artemisinin-treated parasites showed an enrichment in gametogenesis, we did not observe an enrichment for related pathways *per se* in this analysis. We did, however, observe increased gene expression in pathways related to host cell invasion (GO:0044409, symbiont entry into host) and maintaining homeostasis (R-PFA-9646399, aggrephagy) which is essential for parasites growth and development[190]. Interestingly, we also

observed inhibition (decreased expression) of GTP hydrolysis (R-PFA-72706) and formation of free 40S subunits (R-PFA-72689). It is hypothesized that artemisinin mechanisms is related to activation of free radicals in the presence of heme[179], and although the ribosomal 40S subunit is not considered a free radical itself, it can be targeted by free radicals damaging its components and potentially impairing its function[191]. Further investigation could shed light into this as a potential mechanism of artemisinin.

Similar results were obtained for gnf179 and atovaquone. Gnf179 showed enrichment in protein/small molecule transport mechanisms, as well as sphingolipid de novo biosynthesis, consistent with recent studies[177, 178]. Among the enriched pathways in atovaquone, as expected, we found the respiratory electron transport pathway which is dependent of its target Cyt-bc1[174]. Unexpectedly, we observed pathway inhibition of cell signal recognition (R-PFA-1799339) after gnf179 treatment. This could be as an off-target effect due to disruption in the endoplasmic reticulum, as previously shown with ivermectin[192]. For atovaquone, the nonsense mediated decay pathway (R-PFA-975956) was observed with surprisingly high enrichment. Although striking, caution must be taken over this being a direct mechanism of the drug, considering that it is involved in mRNA degradation and could be a result of our sequencing of mostly death parasites. Further investigation into these regulatory changes could provide insights into the parasite's compensatory mechanisms, specific to the chemotype, while in presence of the drug.

4.3 Discussion

In this work, we investigated the use of single-cell sequencing as a method to provide clues about the mechanisms of action for uncharacterized small molecules. We exposed red blood cells infected with *Plasmodium* parasites, and treated them with artemisinin, atovaquone, or gnf179.

Comparison between control and drug-treated parasites showed that these scaffolds result in unique transcriptional profiles relative to the chemotype. We observed gametocytes enrichment for artemisinin treated parasites, and pathway analysis showed enrichment of sphingolipid biosynthesis and electron transport chain for gnf179 and atovaquone, respectively.

Although our results provided some additional clues about the mechanisms of each drug and enrichment of the expected pathways, we also observed problems with this method. For instance, although we obtained a much higher percentage of cells from the first replicate compared to the rest, sequencing mapping showed that those cells were mainly human transcripts rather than *Plasmodium* transcripts, thus had to be filtered. To address this concern, the magnetic enrichment columns (second and third replicate) were used and did provide a much higher coverage of the parasite's transcripts. Additionally, we tried different sequencing depth to evaluate whether it results in more robust results. Indeed, the (third) replicate with the highest depth and magnetic enrichment, showed to be the replicate that yielded the most complete and high-quality cells. Thus, we concluded that parasite enrichment and depth sequencing is essential for this method to work.

Another problem that we faced was the vast difference in parasite population across conditions (i.e., control and drug-treated parasites). Although malaria parasites can be synchronized to ensure having parasites across the asexual stage lifecycle, this process tends to be difficult to achieve and may introduce biases in the drug response evaluation across the cycle. We hypothesized that with single-cell, we could bypass this by detecting expression levels in a variety of different stages at once. In fact, we were able to discern the sub-stages of the asexual lifecycle for all drug-treated parasites, although there were cases where no definitive sub-stage was identified. This could either be because of the drug's effect, or as a defect during cell normalization

and integration steps. Based on our results, evaluation of transcriptomic drug response was only achievable after having a similar number of parasites between the two conditions, thus cell filtering and proper cell integration is also imperative for this approach to work.

Lastly, when inspecting the enriched pathways, the most enriched pathway after atovaquone treatment is related to mRNA degradation, instead of the expected pathway (electron transport). An explanation for this phenomenon could be related to a combination of time of exposure and drug concentration used. Since atovaquone acts as a fast-killing drug, it is likely that parasites were cleared early in their development, or in the process of dying during sequencing pre-processing, ending with residual RNA sequenced unintentionally. Overall, our results show that single-cell sequencing could serve as a method to investigate the mechanisms of action for uncharacterized molecules, although optimization of the experimental conditions (e.g., initial parasite population, time of drug exposure, amount of tested concentration) must be considered prior to considering this approach. Furthermore, this method would only serve as a point of departure for in-depth investigation and validation of the target/mechanisms of action based on the enriched gene/pathways obtained between the set of differentially expressed genes after drug treatment in relation to infected cells.

4.4 Methods

4.4.1 *Plasmodium* Dd2 parasites culture and sequencing

P. falciparum Dd2 asexual blood-stage parasites were grown in human O-positive (O+) whole blood obtained from the BioIVT and San Diego Blood Bank (SDBB) (BioIVT, HUMANRBCPD-0104976) and maintained at 2% hematocrit in RPMI 1640 medium (Gibco). Medium was supplemented with 0.25% Albumax II (Gibco), 2 g/liter sodium bicarbonate (Fisher),

0.1 mM hypoxanthine (Sigma), 25 mM HEPES pH 7.4 (Sigma), and 50 µg/liter gentamicin (Gold Biotechnology). To study the drug effect on the parasites, four flasks were generated, one maintained without drug and the remaining three were dosed with one of atovaquone (50 nM), gnf179 (100 nM), or artemisinin (100 nM). Parasites were incubated at 37ºC and 5% CO2 in the MIC-101 modular incubator chamber. This process was performed in triplicates. After 48 hours of drug exposure, the culture was passed through a column using magnetic cell separation (MACS, Miltenyi Biotec #130-042-901), where infected red blood cells were retained in the column and washed with non-supplemented RPMI medium (except of replicate 1, which was not passed through the magnetic columns but prepared for sequencing after the 48 hours of exposure).

Parasites retained with the magnetic columns were eluted with RPMI non-supplemented medium, concentrated by centrifugation at 1000 g in a 50 µL final volume, and prepared for sequencing following the 10x Genomics protocol (chromium next GEM single-cell 3' HT Reagent, cat #CG000416). Briefly, 27.7 µL infected red blood cells-iRBCs (1,200 iRBCs/µL) were resuspended in 58.7 µL nuclease-free water, loaded into the chromium next GEM chip M for barcoding using standard Illumina primers. cDNA present in each gem was amplified and processed for DNA library construction. These were then sequenced with standard Illumina paired-end constructs using the NovaSeq S4 flow cell, yielding 100-300 (replicate 1 and 2), and 1600M (replicate 3) reads depth.

4.4.2 Bioinformatics workflow

Raw sequences were manually inspected with FastQC[193] for quality check. A custom reference genome was created with CellRanger[194] *mkref* tool, using the human (GRch38.p13) and *Plasmodium* (PfDd2 v62, PlasmoDB[131]) genomes. Raw sequences were then mapped to the

custom genome and transcripts counted using CellRanger count function. Filtered barcoded features were then analyzed using Seurat's functions[181]. Features were first opened with *Read10X* function, and *Plasmodium* features where extracted. From these, protein coding gene transcripts were filtered and used to create a Seurat object.

For each sample (control, artemisinin-, gnf179-, and atovaquone-treated cells), replicate objects were merged and a total of 15,000 cells where randomly selected. Cells where then normalized, highly variable features were identified and scaled, dimensional reduction was performed (PCA), and batch-correction was performed using Harmony. Cell clusters were obtained using the Harmony correction. To contrast cell conditions, control cells were merged with one of the treated samples (artemisinin, gnf179 and atovaquone). A standard pipeline (highly variable features were detected, scaled and dimensional reduction) was performed. Cells where then batch corrected using Harmony and cells' layers where integrated to find the clusters. For each object, cells with less than 50 genes were filtered and clusters with less than 1,000 cells were removed. Hypervariable gene families (*rifin*, *stevor*, *var*) were filtered from analyses.

4.4.3 Cell marker identification

Gene markers for each integrated condition (control vs drug-treated) were identified using Seurat's *FindAllMarkers* function, keeping genes with at least 0.25 log fold-change. These were used to assign cell cluster ID's based on known markers or similar expression from the Malaria Cell Atlas dataset[140] (**Table 4.1**). Shortly, the list of potential gene markers for each (unassigned) cluster was sorted by *p*-adjusted value (lowest to highest) and percentage of cells (highest to lowest) expressing that gene in one cluster (pct.1) compared to the percentage in all the remaining clusters (pct.2). In cases a known gene marker (i.e., literature established markers across the

Plasmodium asexual lifecycle) was present among the top 20 genes based on the sorting, the associated cluster in pct.1 was determined as the marker for said cluster and confirmed using the *FeaturePlot* function in Seurat. In cases where no known marker was found, starting from the first gene in the sorting, if the difference between pct.1 and pct.2 was of at least 60%, the potential gene marker was visualized in the dimensional graph using *FeaturePlot* function in Seurat; if confirmed to be specific to one stage (e.g., highly expressed in one cluster compared to the rest), the gene was then queried using the Malaria Cell Atlas to label the asexual sub-stage expressing said gene.

4.4.4 Differential gene expression

After cluster determination for each of the Seurat's objects (ctr-atovaquone, ctr-artemisinin, and ctr-gnf179), we used the metadata from each object to add a column (clust.cond) using a combination between the cluster number and the condition (control or drug treated) and was used to re-identify the identifiers from each Seurat object. Next, for each cluster in the object we used the *FindMarkers* function using as ident.1 the treated cells and ident.2 the control cells, as previously defined in clust.cond. Wilcox test was used to calculate the degree of difference between each condition, with default settings for the remaining options. The same process was used to compare clusters between each other.

4.4.5 Finding 3D7 orthologs

Each Dd2 identifier was taken and queried in PlasmoDB database gene strategy function. A complete query from the list was performed with the latest *P. falciparum* Dd2 release (accessed on April 2025). Next, a download with customizable data was performed (tab format), to download the list Orthologs and Paralogs data within VEuPathDB[131] database. This query resulted in the list

of Dd2 genes, with the corresponding ortholog in other *Plasmodium* strain or species. Hits for *P. falciparum* 3D7 were extracted from this list, including the protein product for the matched ortholog; there were cases where two or more orthologs were found for one Dd2 gene.

4.4.6 Pathway enrichment analysis

Taking the list of 3D7 orthologs of interest, we split two groups of genes, one for upregulated genes (having gene expression greater than zero), and downregulated genes (gene expression less than zero). We also filtered genes having a *p*-adjusted value ≤ 0.05. Two queries were performed; one for expression values of at least 1–fold (both up- and down-regulated) and the other for at least 3–fold (both up- and down-regulated). Each list was queried in Metascape[188] online server, pasting the list of genes, selecting queries for *P. falciparum* genes in both species input and analysis species. An express analysis was performed for each query. Since we obtained some pathways from human inferences not pertinent to *Plasmodium*, we filtered those out from the results.

4.5 Acknowledgements

Chapter 4 contains unpublished material co-authored with Elizabeth A. Winzeler. The dissertation author was the primary author of this chapter.

Figure 4.1 Experimental design and processing of *P. falciparum* asexual parasites

(A) Experimental design workflow. Steps followed to study drug treatment in *Plasmodium falciparum* Dd2 parasites. Three biological replicates were performed; human red blood cells were infected with Dd2 parasites, one was left untreated (control) and three separated iRBC were treated with atovaquone, artemisinin and gnf179 at 10x IC_{50} concentrations. After 48 hours post infection, parasites were prepared for 10x Chromium single-cell protocol and sequenced using NovaSeq S4 flow cell (NovaSeq 6000 system). A custom bioinformatic workflow was followed, using a combined human and *Plasmodium* genome for alignment and transcript quantification. Downstream analyses were performed to *Plasmodium* cells only. **(B) Number of raw cells obtained and downsampled across samples.** Histogram contrasting the total number of cells acquired across control (blue), artemisinin (purple), gnf179 (pink) and atovaquone (green) samples. Mean number of raw cells are shown for each replicate of control (n = 124,904), artemisinin (n = 29,174), gnf179 (n = 93,705) and atovaquone (n = 91,108.3) samples. Mean number of cells per replicate are shown relative to the right axis, for control (n = 4,724.33), artemisinin (n = 4,751), gnf179 (n = 3,000.67) and atovaquone (n = 3,068), after randomly selecting 15,000 cells per sample. Error bars for mean values are included for each sample. **(C) Single-cell integration and correction workflow.** Pre-processing steps followed to integrate control cells and drug-treated cells. The package tool Seurat[181] was used for the processing and analysis of cells. Replicates for each sample were integrated as a single object (control R1-3, atovaquone R1-3, artemisinin R1-3 and gnf179 R1-3) using the merge option, respectively. Objects were subject to a normal single-cell pipeline correcting using Harmony by replicate (normalization, cell cluster obtention, batch correction). The corrected objects, were integrated using Seurat's merge function, having x as the control and y as the drug-treated (artemisinin, atovaquone, or gnf-179), and subject to a standard single-cell pipeline and batch correcting according to the infected or treated condition. Genes in hypervariable regions of the genome and outlier cells were filtered. **(D) Cell expression profile comparison.** High-quality cells were visually confirmed by plotting the total number of counts (X axis) versus number of genes (Y axis) for each integrated sample. Control cells are shown in blue, artemisinin-treated cells in purple, gnf179-treated cells are in pink and atovaquone-treated cells are shown in green.

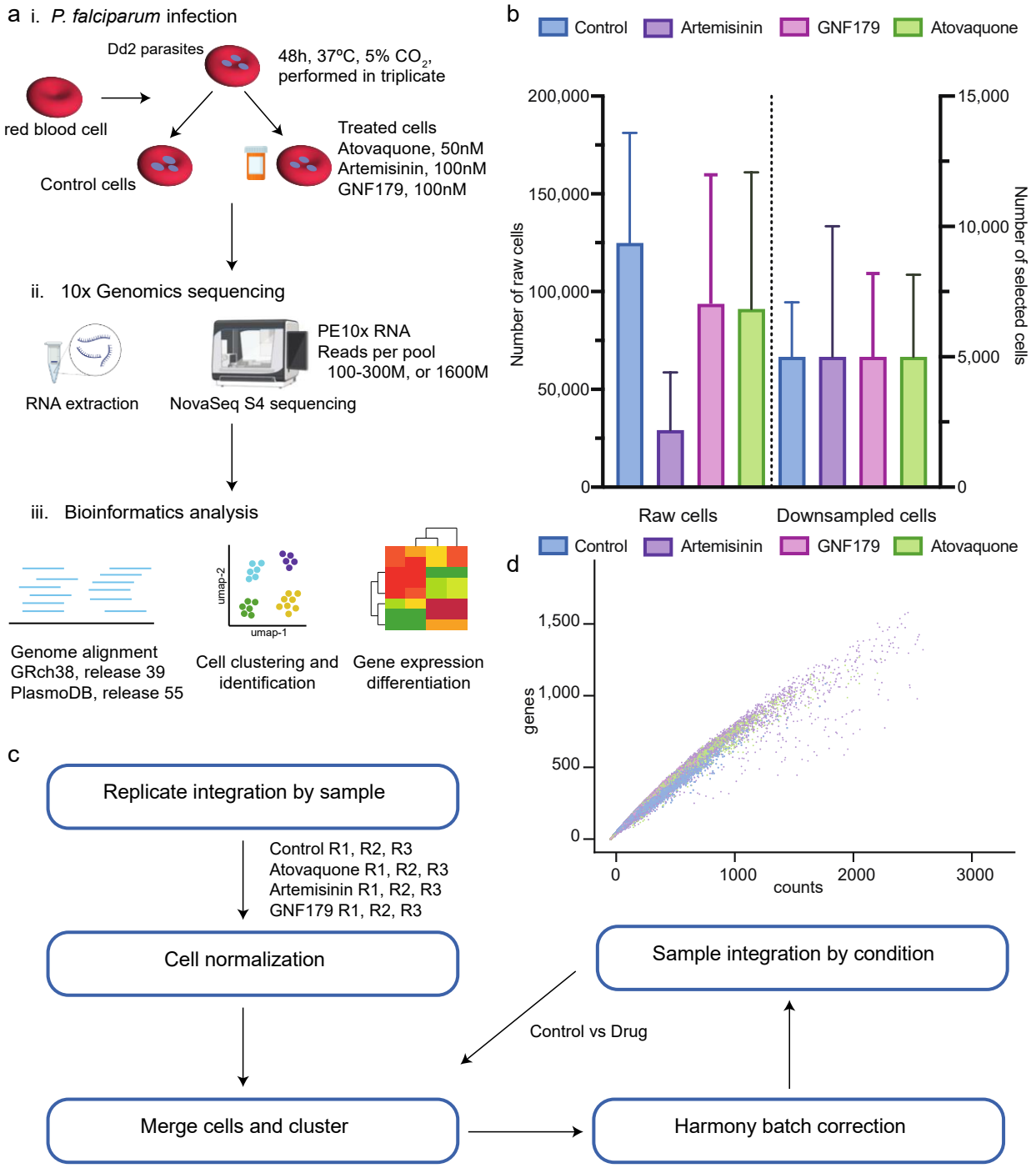

a
i. P. falciparum infection
Dd2 parasites
48h, 37ºC, 5% CO2,
performed in triplicate
red blood cell
Treated cells
Atovaquone, 50nM
Artemisinin, 100nM
GNF179, 100nM
Control cells
ii. 10x Genomics sequencing
PE10x RNA
Reads per pool
100-300M, or 1600M
RNA extraction
NovaSeq S4 sequencing
iii. Bioinformatics analysis
umap-2
umap-1
Genome alignment
GRch38, release 39
PlasmoDB, release 55
Cell clustering and
identification
Gene expression
differentiation
b
Control
Artemisinin
GNF179
Atovaquone
200,000
150,000
100,000
50,000
0
Number of raw cells
15,000
10,000
5,000
0
Number of selected cells
Raw cells
Downsampled cells
c
Replicate integration by sample
Control R1, R2, R3
Atovaquone R1, R2, R3
Artemisinin R1, R2, R3
GNF179 R1, R2, R3
Cell normalization
Sample integration by condition
Control vs Drug
Merge cells and cluster
Harmony batch correction
d
Control
Artemisinin
GNF179
Atovaquone
1,500
1,000
500
0
genes
0
1000
2000
3000
counts

Figure 4.2 Transcriptional profiling of asexual *Plasmodium* parasites relative to after drug treatment

(A) Cell clustering profile of parasites after artemisinin, gnf179 and atovaquone treatment. Dimensional reduction of high-quality cells relative to after drug treatment were visualized using Seurat's DimPlot function. Clusters were identified using known gene markers and are shown next to each cluster. Cases with no clear indication of the parasite's stage were marked as "NA". **(B) Fold-change transcription changes for 281 unique genes after artemisinin, 159 genes for atovaquone and 132 genes for gnf-179 exposure in contrast to non-treated parasites.** Genes with at least $\log_2$ fold of -3 or 3 after differential gene expression analysis were taken from **Supplemental Table 4.1**. Samples (artemisinin, atovaquone, and gnf179) were clustered hierarchically based on gene-expression profiles. Expression fold change are represented from low (dark red) to high (dark blue); no change is shown in light orange. Genes with an altered expression profile for at least two treatments are labeled on the right panel, and cases with opposite profiles across treatments are indicated by arrows on the left panel. **(C) Overlap of 475 genes with $\log_2$ fold of at least ±3.** Overlap of genes with at least $\log_2$ fold change of -3 or 3 across treatments. A total of 63 genes were transcriptionally different across two treatments and only 17 genes were transcriptionally altered across all.

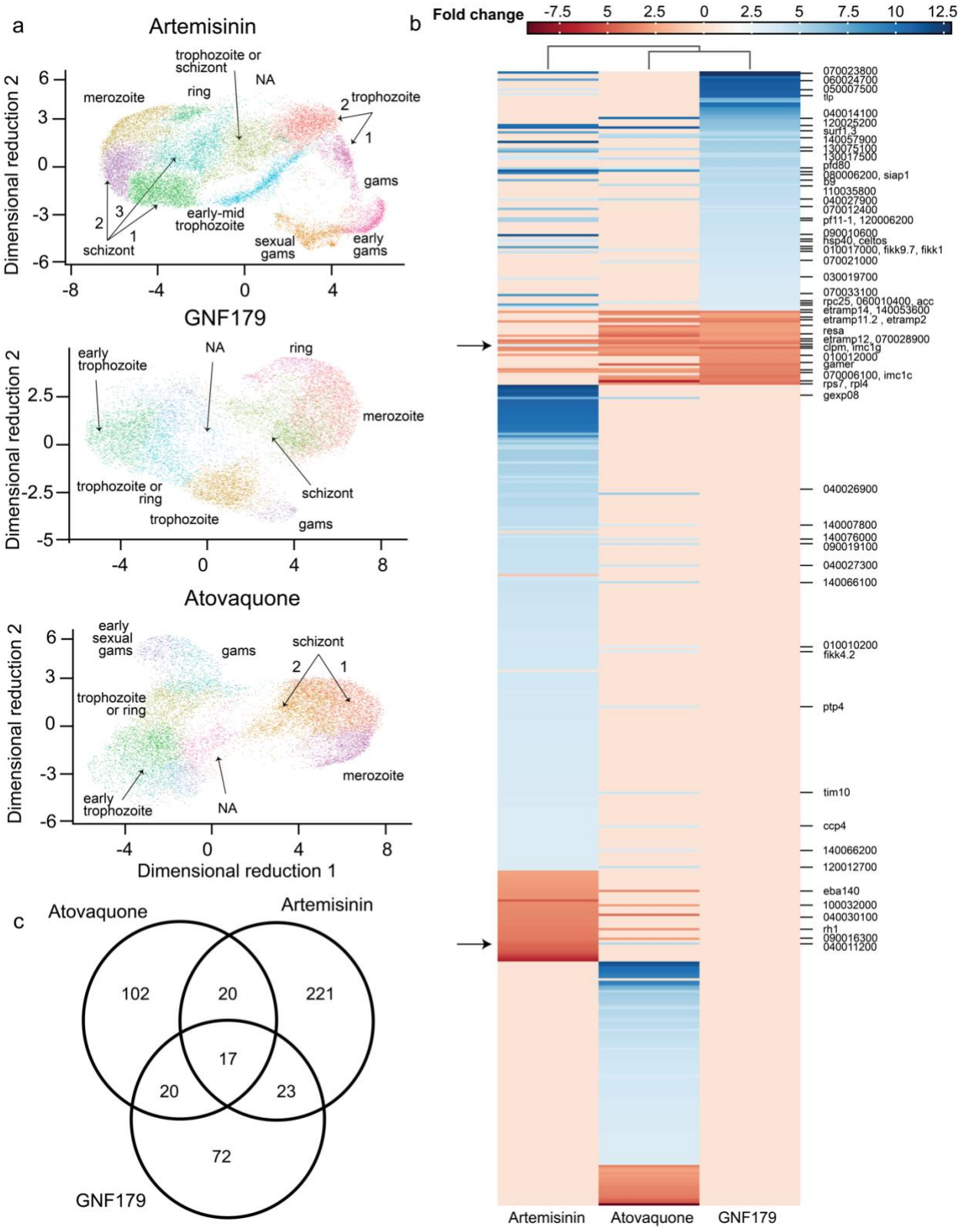

a
Artemisinin
trophozoite or schizont
NA
merozoite
ring
2
trophozoite
1
gams
3
early-mid trophozoite
2
1
schizont
sexual gams
early gams
Dimensional reduction 2
GNF179
early trophozoite
NA
ring
merozoite
trophozoite or ring
schizont
trophozoite
gams
Atovaquone
early sexual gams
gams
schizont
2
1
trophozoite or ring
merozoite
early trophozoite
NA
Dimensional reduction 1
b
Fold change
-7.5
5
2.5
0
2.5
5
7.5
10
12.5
070023800
060024700
050007500
tlp
040014100
120025200
surf1.3
140057900
130075100
130017500
pfd80
080006200, siap1
b9
110035800
040027900
070012400
pf11-1, 120006200
090010600
hsp40_celtos
010017000, fikk9.7, fikk1
070021000
030019700
070033100
rpc25, 060010400, acc
etramp14, 140053600
etramp11.2 , etramp2
resa
etramp12, 070028900
clpm, imc1g
010012000
gamer
070006100, imc1c
rps7, rpl4
gexp08
040026900
140007800
140076000
090019100
040027300
140066100
010010200
fikk4.2
ptp4
tim10
ccp4
140066200
120012700
eba140
100032000
040030100
rh1
090016300
040011200
Artemisinin
Atovaquone
GNF179
c
Atovaquone
Artemisinin
102
20
221
17
20
23
72
GNF179

Figure 4.3 Enriched pathways relative to after drug treatment

(A) Total number of Dd2 upregulated (*n* = 2,808) and downregulated (*n* = 1,339) genes, per treatment. Breakdown of differentially expressed genes having a fold-change ≤ -1 (downregulated) or fold-change ≥ 1 (upregulated) as determined in **Supplemental Table 4.1**, with significance -Log_{10}(*p*-value) of ≤ 0.05. Exact number of genes having an increased or decreased transcription (*Y* axis) after drug treatment relative to control cells for artemisinin (up, *n* = 1,486; down *n* = 249), gnf179 (up, *n* = 353; down *n* = 893), and atovaquone (up, *n* = 969; down *n* = 197) are shown. Transcriptional changes are denoted by color, having increased transcription in blue, and decreased transcription in purple. **(B) Upset plot showing 3D7 orthologs for differentially expressed genes per treatment and overlap among treatments.** Bottom panel shows distribution of 3D7 ortholog matches (*n* = 2,603, not unique) for the 4,147 Dd2 differentially expressed genes. 44 Dd2 genes are not included since no ortholog was found. *X* axis shows the number of orthologs per transcriptional change set (upregulated in blue, downregulated in purple). Top panel shows the number of ortholog intersection for both transcriptional groups (i.e., same ortholog shared by one or more treatments in the upregulated or downregulated set). Number of upregulated orthologs are shown in the left *Y* axis (blue), and number of downregulated orthologs are shown in the right *Y* axis (purple). Intersection groups (i.e., drug treatment) are indicated in the middle panel, with black dots in cases where genes are unique to the drug treatment (e.g., atovaquone, artemisinin and gnf179), and with connecting lines in the case of genes shared by two or more drugs (e.g., gnf179 and artemisinin). **(C) Enriched pathways for artemisinin, gnf179 and atovaquone, relative to after drug treatment.** Barplot showing the enriched pathways (-Log_{10} p-value > 1) for each drug treatment, as determined by Metascape[188] using the *P. falciparum* 3D7 genes orthologs. Pathway memberships are shown above (or below) each bar, having in blue pathways that increased in the cell and in purple pathways that decreased RNA production in the cell. *P*-values for each set are shown in the *Y*-axis, each starting at zero and increasing positively, to facilitate the visualization between up- down-regulated pathways.

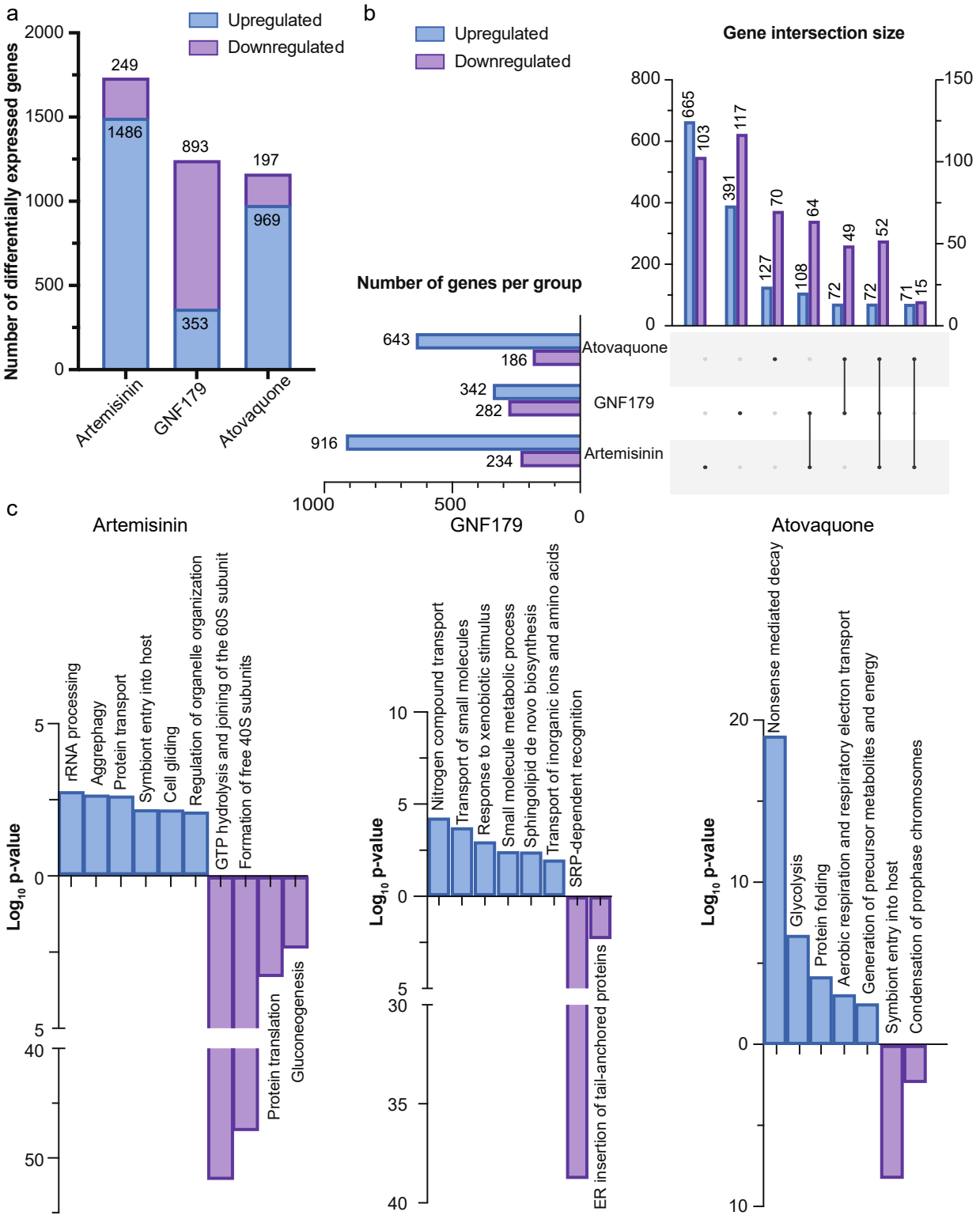

a
Upregulated
Downregulated
Number of differentially expressed genes
2000
1500
1000
500
0
249
1486
893
353
197
969
Artemisinin
GNF179
Atovaquone
b
Upregulated
Downregulated
Number of genes per group
643
186
342
282
916
234
1000
500
0
Atovaquone
GNF179
Artemisinin
Gene intersection size
800
600
400
200
0
150
100
50
0
665
103
117
391
70
127
64
108
49
72
52
72
71
15
c
Artemisinin
Log10 p-value
5
0
5
40
50
rRNA processing
Aggrephagy
Protein transport
Symbiont entry into host
Cell gliding
Regulation of organelle organization
GTP hydrolysis and joining of the 60S subunit
Formation of free 40S subunits
Protein translation
Gluconeogenesis
GNF179
Log10 p-value
10
5
0
5
30
35
40
Nitrogen compound transport
Transport of small molecules
Response to xenobiotic stimulus
Small molecule metabolic process
Sphingolipid de novo biosynthesis
Transport of inorganic ions and amino acids
SRP-dependent recognition
ER insertion of tail-anchored proteins
Atovaquone
Log10 p-value
20
10
0
10
Nonsense mediated decay
Glycolysis
Protein folding
Aerobic respiration and respiratory electron transport
Generation of precursor metabolites and energy
Symbiont entry into host
Condensation of prophase chromosomes

Table 4.1 Gene markers used for asexual stage cell type identification.

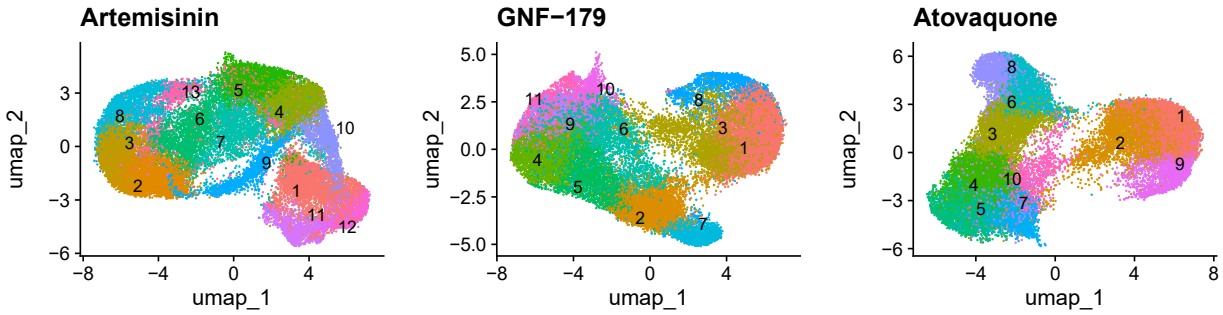
List of markers used for cluster target identification. Source showing validation of the gene as the lifecycle development stage marker is included, or relative gene expression as determined in the Malaria Cell Atlas study[140].

Lifecycle stage	Known marker	Malaria Cell Atlas expression
Ring	PfDd2_010006000 (*resa*)[195], PfDd2_110051800 (*resa*)[195], PfDd2_020005400 (*etramp2*) [186]	
Trophozoite	PfDd2_110051600 (*pf332*)[196], PfDd2_010016500 (*garp*)[197]	PfDd2_080033000 (*pdi8,* PBANKA_0702800)
Ring or trophozoite		PfDd2_100017800 (*hgprt*, PBANKA_1210800)
Early trophozoite	PfDd2_010016500 (*garp*)[198]	
Early/mid trophozoite	PfDd2_110051600 (*pf332*)[199]	
Schizont		PfDd2_140015000 (*rap1*, PBANKA_1032100) PfDd2_060032800 (*rama*, PBANKA_0804500), PfDd2_120057300 (*ron3*, PBANKA_1464900)
Merozoite	PfDd2_070034500 (*eba175*)[200], PfDd2_090035400 (*msp1*)[201], PfDd2_130040900 (*msp7*)[201], PfDd2_120033600 (*msp9*)[201]	
Gametocytes	PfDd2_100043400 (*pf11_1*)[202]	
Early gametocytes	PfDd2_130007700 (*g27/25*)[203]	
Sexual gametocytes	PfDd2_040011600 (*pfs16*)[204], PfDd2_120021200 (*mdv1*)[205]	

Table 4.2 Gene differences between schizont clusters in artemisinin and atovaquone treatment, and trophozoite clusters in artemisinin treatment.

Gene expression differentiation was calculated for both treatments, utilizing Seurat's *FindMarkers* function. Schizont and Trophozoite parasites were compared according to **Figure 4.2A** and are indicated in "Cluster" column. Fold changes were calculated based on Wilcox test, and artemisinin schizont stage was performed comparing one stage versus the other two combined. *P*-values were calculated according to *t*-test and adjusted *p*-values are Bonferroni corrected.

Treatment	Cluster	Gene	Log_2 fold change	P-value
Atovaquone	Schizont 1 vs 2	*pfeba-181* (erythrocyte binding antigen-181)	1.15	$1.04x10^{-128}$
		pfama1 (apical membrane antigen 1)	1.21	$1.23x10^{-95}$
		pfmtrap (merozoite TRAP-like protein)	1.06	$7.81x10^{-91}$
		pfetramp2 (early transcribed membrane protein 2)	-1.11	$7.74x10^{-5}$
Artemisinin	Trophozoite 1 vs 2	*pfg27/25* (gamete antigen 27/25)	1.77	$3.78x10^{-162}$
		pfd80 (*Plasmodium* exported protein PHISTb)	1.16	$2.67x10^{-137}$
		pfetramp5 (early transcribed membrane protein 5)	-2.16	$1.72x10^{-89}$
		pfhgprt (hypoxanthine-guanine phosphoribosyltransferase)	-2.60	$2.31x10^{-81}$
	Schizont 1 vs 2, 3	*pfapi-ap2* (transcription factor with AP2 domain)	0.36	$1.28x10^{-13}$
	Schizont 2 vs 1, 3	*pfeba-175* (erythrocyte binding antigen 175)	1.16	$8.18 x10^{-249}$
	Schizont 3 vs 1, 2	*pfp2* (60S acidic ribosomal protein P2)	0.36	$1.28x10^{-05}$



Supplemental Figure 4.1 Clustering for artemisinin-, gnf179-, and atovaquone-treated cells versus non-treated cells.

Plots showing artemisinin, gnf179 and atovaquone dimensional reduction after cell integration and batch correction steps. Clusters were determined using a standard Seurat pipeline (see **methods**), with resolution of 0.8. Cluster numbers and colors are shown with default settings and represent the total of distinct clusters obtained after using *FindNeighbors* and *FindClusters* Seurat's functions. Overlapping clusters are denoted by cells (color 1 dots) mainly observed overlaying on another cluster (color 2 dots): gnf179 cluster 9, 10 and 11, and atovaquone cluster 5 and 7.

Chapter 5

Conclusions

Neglected tropical diseases (NTD) –caused by viruses, bacteria, parasites, fungi and toxins– affects 1 billion people worldwide according to the World Health Organization. Although devastating associated diseases have high prevalence among impoverished communities, funding resources available are limited and they attract less commercial interest in contrast to other diseases such as cancer or diabetes. Over the past decade, much progress in prevention and control has been made, mainly in drug development through target discovery and validation. In particular, phenotypic screening has gained popularity since large, diverse compound libraries can be tested in a high-throughput fashion without requiring a priori knowledge of targets or mechanisms of action. However, the methods tend to be time- and resource-consuming, limited by phenotypic assay conditions, and there is a rise of drug resistance to current phenotypically derived therapeutic inhibitors. Thus, in this dissertation we use computational approaches to better understand small molecules and identify novel therapeutic drug targets, focusing on *Plasmodium* parasites as a model due to the vast known phenotypic information and genomic knowledge compared to other neglected diseases.

In Chapter 1, we focused on facilitating the exploration of large chemical and biological datasets, aiming to provide a better understanding of chemical structures and their impact on a biological system, as well as promptly identify small molecules with well-established mechanism or known drug resistance liabilities in the context of drug discovery. We thus constructed a tool, called CACTI, that utilizing data mining and chemoinformatic techniques would automate the tedious molecule-by-molecule searches providing a comprehensive knowledge report and potential drug-target in an unbiased way. We assessed CACTI's performance by querying the Pathogen Box library, consisting of 400 molecules with diverse scaffolds and active against a variety of neglected tropical diseases (including malaria and toxoplasma). The library contained

26 controls that were removed from the results. The remaining 374 compounds were queried to find synonym names (i.e. molecule identifiers), literature citations, close analogs, and potential drug-target clusters. We obtained a total of 4,315, 30,425, 12,383 and 77, respectively, that were unreported prior to the library publication. Further inspection into these findings showed that most of the compounds grouped among the 77 clusters are similar to known inhibitors, including dihydroorotate dehydrogenase and cytochrome B. This tool also has liabilities; we rely on similar scaffolds having same target assumptions, veracity/strength of automated annotations are not assessed, and the scaffold comparison approach is not suitable for larger molecules including natural products. Nevertheless, this tool provides a starting point for target validation and accelerates drug discovery with the rapid evaluation of large chemical libraries.

In Chapter 2 we further explore small molecules liabilities, specifically in drug resistance, to reveal features of resistance mutations in *Plasmodium falciparum*, the malaria-causing parasite. A collection of 724 whole-genome *P. falciparum* sequences, containing genomes from in vitro evolved resistant clones and their isogenic parent, were analyzed collectively. An in vitro evolution and whole genome sequencing analysis revealed 5,560 SNVs and indels and 1,168 CNVs, when comparing the resistant clone versus their corresponding parent. From these, we observed 128 mutated genes appearing at rates greater than expected, and co-occurrence of some genes such as *pfapi-ap2* (with *pfmdr1*), suggesting this gene may play a role in contributing to multidrug resistance in *Plasmodium*. Altogether, these results form the basis for predicting which genetic alleles are likely to confer drug resistance in malaria parasites and other pathogens, setting the foundation for machine learning predictive approaches and improving available drug-target prediction tools.

In Chapter 3, we identified a list of novel drug targets differ in mechanism of action and resistance liabilities to current therapeutic small molecules. Through a comprehensive extraction and integration of genomic and chemogenomic databases, we systematically assessed the *Plasmodium* genome and identified 867 candidate protein targets with evidence of small-molecule binding and blood-stage essentiality. Of these, 540 proteins showed strong essentiality evidence including *PfDHODH*, and *PfDHFR*, which served as validation of our approach; we also found proteins that lack inhibitors that have progressed to clinical trials such as *PfBDP1*. The list of 540 candidate targets were further narrowed down following expert review and rubric-based scoring, to select proteins with highly attractive characteristics to known drug targets such as selectivity, structural information, and assay developability. This exercise yielded 27 high-priority antimalarial target candidates, many of which are considered novel (i.e. limited prior characterization, especially as an antimalarial target). Here, we provide a list of highly attractive antimalarial targets for validation, and we provide a generalizable framework for systematically evaluating and prioritizing of novel pathogenic disease targets.

Lastly, in Chapter 4 use single-cell sequencing as a method to identify the mechanisms of action for different chemotypes, particularly (uncharacterized) molecules that have failed through other approaches. By contrasting transcriptomic profiles of human red blood cells infected with *P. falciparum* and drug-treated cells, we found that expression patterns for atovaquone, artemisnin, and gnf179 are unique to each chemotype, with artemisinin showing enrichment on gametocyte stages while atovaquone did not exhibited a ring stage. Furthermore, we observed genes differentially expressed (DEG) between control and drug-treated parasites, resulting in 3,172 DEG for artemisinin, 1,488 DEG genes for gnf179, and 2,239 DEG genes after atovaquone treatment. A pathway analysis for a subset of these upregulated genes showed enrichment of the electron

transport pathway in atovaquone and sphingolipid de novo biosynthesis pathway in gnf179, as well as additional hints for artemisinin not explored in the field as of now. Although promising, further optimization and validations are needed to confidently support these finding and the use of single-cell to characterize molecules in malaria and other pathogenic diseases.

REFERENCES

1 Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K. & Yamanishi, Y. Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers. *Bioinformatics* **28**, i487-i494 (2012). https://doi.org:10.1093/bioinformatics/bts412

2 Schenone, M., Dancik, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* **9**, 232-240 (2013). https://doi.org:10.1038/nchembio.1199

3 Zhou, W., Wang, Y., Lu, A. & Zhang, G. Systems pharmacology in small molecular drug discovery. *Int J Mol Sci* **17**, 246 (2016). https://doi.org:10.3390/ijms17020246

4 Rao, S. P. S., Manjunatha, U. H., Mikolajczak, S., Ashigbie, P. G. & Diagana, T. T. Drug discovery for parasitic diseases: powered by technology, enabled by pharmacology, informed by clinical science. *Trends Parasitol* **39**, 260-271 (2023). https://doi.org:10.1016/j.pt.2023.01.010

5 Hovlid, M. L. & Winzeler, E. A. Phenotypic screens in antimalarial drug discovery. *Trends Parasitol* **32**, 697-707 (2016). https://doi.org:10.1016/j.pt.2016.04.014

6 Sink, R., Gobec, S., Pecar, S. & Zega, A. False positives in the early stages of drug discovery. *Curr Med Chem* **17**, 4231-4255 (2010). https://doi.org:10.2174/092986710793348545

7 Chaniad, P., Mungthin, M., Payaka, A., Viriyavejakul, P. & Punsawad, C. Antimalarial properties and molecular docking analysis of compounds from *Dioscorea bulbifera L.* as new antimalarial agent candidates. *BMC Complement Med Ther* **21**, 144 (2021). https://doi.org:10.1186/s12906-021-03317-y

8 Owoloye, A., Enejoh, O. A., Akanbi, O. M. & Bankole, O. M. Molecular docking analysis of *Plasmodium falciparum* dihydroorotate dehydrogenase towards the design of effective inhibitors. *Bioinformation* **16**, 672-678 (2020). https://doi.org:10.6026/97320630016672

9 Owoloye, A. J., Ligali, F. C., Enejoh, O. A., Musa, A. Z., Aina, O., Idowu, E. T. & Oyebola, K. M. Molecular docking, simulation and binding free energy analysis of small molecules

as *Pf*HT1 inhibitors. *PLoS One* **17**, e0268269 (2022). https://doi.org:10.1371/journal.pone.0268269

10 Singh, N., Chaput, L. & Villoutreix, B. O. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Brief Bioinform* **22**, 1790-1818 (2021). https://doi.org:10.1093/bib/bbaa034

11 de Sousa, A. C. C., Combrinck, J. M., Maepa, K. & Egan, T. J. Virtual screening as a tool to discover new beta-haematin inhibitors with activity against malaria parasites. *Sci Rep* **10**, 3374 (2020). https://doi.org:10.1038/s41598-020-60221-0

12 Uddin, A., Gupta, S., Mohammad, T., Shahi, D., Hussain, A., Alajmi, M. F., El-Seedi, H. R., Hassan, I., Singh, S. & Abid, M. Target-based virtual screening of natural compounds identifies a potent antimalarial with selective falcipain-2 inhibitory activity. *Front Pharmacol* **13**, 850176 (2022). https://doi.org:10.3389/fphar.2022.850176

13 Godara, P., Reddy, K. S., Sahu, W., Naik, B., Srivastava, V., Das, R., Mahor, A., Kumar, P., Giri, R., Anirudh, J., Tak, H., Banavath, H. N., Bhatt, T. K., Goyal, A. K. & Prusty, D. Structure-based virtual screening against multiple *Plasmodium falciparum* kinases reveals antimalarial compounds. *Mol Divers* **28**, 3661-3681 (2024). https://doi.org:10.1007/s11030-023-10770-z

14 Wang, L., Ma, C., Wipf, P., Liu, H., Su, W. & Xie, X. Q. TargetHunter: an in silico target identification tool for predicting therapeutic potential of small organic molecules based on chemogenomic database. *AAPS J* **15**, 395-406 (2013). https://doi.org:10.1208/s12248-012-9449-z

15 Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. & Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100-1107 (2012). https://doi.org:10.1093/nar/gkr777

16 Backman, T. W., Cao, Y. & Girke, T. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res* **39**, W486-491 (2011). https://doi.org:10.1093/nar/gkr320

17 Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J. & Bryant, S. H. PubChem substance and

compound databases. *Nucleic Acids Res* **44**, D1202-1213 (2016). https://doi.org:10.1093/nar/gkv951

18 Jimenes-Vargas, K., Pazos, A., Munteanu, C. R., Perez-Castillo, Y. & Tejera, E. Prediction of compound-target interaction using several artificial intelligence algorithms and comparison with a consensus-based strategy. *J Cheminform* **16**, 27 (2024). https://doi.org:10.1186/s13321-024-00816-1

19 Gregson, A. & Plowe, C. V. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacol Rev* **57**, 117-145 (2005). https://doi.org:10.1124/pr.57.1.4

20 Yang, S. Q., Zhang, L. X., Ge, Y. J., Zhang, J. W., Hu, J. X., Shen, C. Y., Lu, A. P., Hou, T. J. & Cao, D. S. In-silico target prediction by ensemble chemogenomic model based on multi-scale information of chemical structures and protein sequences. *J Cheminform* **15**, 48 (2023). https://doi.org:10.1186/s13321-023-00720-0

21 Ji, K. Y., Liu, C., Liu, Z. Q., Deng, Y. F., Hou, T. J. & Cao, D. S. Comprehensive assessment of nine target prediction web services: which should we choose for target fishing? *Brief Bioinform* **24** (2023). https://doi.org:10.1093/bib/bbad014

22 Chen, X., Liu, M. & Gilson, M. K. BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* **4**, 719-725 (2001). https://doi.org:10.2174/1386207013330670

23 Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A. & Rapp, B. A. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **28**, 10-14 (2000). https://doi.org:10.1093/nar/28.1.10

24 Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S. A., Pettersson, J., Goncharoff, N., Hersey, A. & Overington, J. P. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res* **44**, D1220-1228 (2016). https://doi.org:10.1093/nar/gkv1253

25 Shaikh, F., Tai, H. K., Desai, N. & Siu, S. W. I. LigTMap: ligand and structure-based target identification and activity prediction for small molecular compounds. *J Cheminform* **13**, 44 (2021). https://doi.org:10.1186/s13321-021-00523-1

26 Nickel, J., Gohlke, B. O., Erehman, J., Banerjee, P., Rong, W. W., Goede, A., Dunkel, M. & Preissner, R. SuperPred: update on drug classification and target prediction. *Nucleic Acids Res* **42**, W26-31 (2014). https://doi.org:10.1093/nar/gku477

27 Awale, M. & Reymond, J. L. Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. *J Chem Inf Model* **59**, 10-17 (2019). https://doi.org:10.1021/acs.jcim.8b00524

28 Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L. & Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **43**, W612-620 (2015). https://doi.org:10.1093/nar/gkv352

29 McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P. & Lopez, R. Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res* **41**, W597-600 (2013). https://doi.org:10.1093/nar/gkt376

30 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput* **28**, 31-36 (1988).

31 RDKit. RDKit: open-source cheminformatics.

32 Bajusz, D., Racz, A. & Heberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 20 (2015). https://doi.org:10.1186/s13321-015-0069-3

33 Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D. & Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J Med Chem* **39**, 3049-3059 (1996). https://doi.org:10.1021/jm960290n

34 Canham, S. M., Wang, Y., Cornett, A., Auld, D. S., Baeschlin, D. K., Patoor, M., Skaanderup, P. R., Honda, A., Llamas, L., Wendel, G., Mapa, F. A., Aspesi, P., Jr., Labbe-Giguere, N., Gamber, G. G., Palacios, D. S., Schuffenhauer, A., Deng, Z., Nigsch, F., Frederiksen, M., Bushell, S. M., Rothman, D., Jain, R. K., Hemmerle, H., Briner, K., Porter, J. A., Tallarico, J. A. & Jenkins, J. L. Systematic chemogenetic library assembly. *Cell Chem Biol* **27**, 1124-1129 (2020). https://doi.org:10.1016/j.chembiol.2020.07.004

35 Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Spedding, M. & Davies, J. A. The IUPHAR/BPS guide to pharmacology in 2024. *Nucleic Acids Res* **52**, D1438-D1449 (2024). https://doi.org:10.1093/nar/gkad944

36 Administration, F. D. Drugs@FDA: FDA-approved drugs.

37 Mugumbate, G. & Overington, J. P. The relationship between target-class and the physicochemical properties of antibacterial drugs. *Bioorg Med Chem* **23**, 5218-5224 (2015). https://doi.org:10.1016/j.bmc.2015.04.063

38 Raimondi, M. V., Randazzo, O., La Franca, M., Barone, G., Vignoni, E., Rossi, D. & Collina, S. DHFR inhibitors: reading the past for discovering novel anticancer agents. *Molecules* **24** (2019). https://doi.org:10.3390/molecules24061140

39 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003). https://doi.org:10.1101/gr.1239303

40 Venture, M. f. M. Pathogen Box, <https://www.mmv.org/mmv-open/pathogen-box/about-pathogen-box> (2016).

41 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3-26 (2001). https://doi.org:10.1016/s0169-409x(00)00129-0

42 Wong, W., Bai, X. C., Sleebs, B. E., Triglia, T., Brown, A., Thompson, J. K., Jackson, K. E., Hanssen, E., Marapana, D. S., Fernandez, I. S., Ralph, S. A., Cowman, A. F., Scheres, S. H. W. & Baum, J. Mefloquine targets the *Plasmodium falciparum* 80S ribosome to inhibit protein synthesis. *Nat Microbiol* **2**, 17031 (2017). https://doi.org:10.1038/nmicrobiol.2017.31

43 Cheng, J., Ma, X., Krausz, K. W., Idle, J. R. & Gonzalez, F. J. Rifampicin-activated human pregnane X receptor and CYP3A4 induction enhance acetaminophen-induced toxicity. *Drug Metab Dispos* **37**, 1611-1621 (2009). https://doi.org:10.1124/dmd.109.027565

44 Tiash, S., Saunders, J., Hart, C. J. S., Ryan, J. H., Riches, A. G. & Skinner-Adams, T. S. An image-based Pathogen Box screen identifies new compounds with anti-Giardia activity and highlights the importance of assay choice in phenotypic drug discovery. *Int J Parasitol Drugs Drug Resist* **12**, 60-67 (2020). https://doi.org:10.1016/j.ijpddr.2020.03.002

45 Dennis, A. S. M., Rosling, J. E. O., Lehane, A. M. & Kirk, K. Diverse antimalarials from whole-cell phenotypic screens disrupt malaria parasite ion and volume homeostasis. *Sci Rep* **8**, 8795 (2018). https://doi.org:10.1038/s41598-018-26819-1

46 Veale, C. G. L. Unpacking the Pathogen Box-an open source tool for fighting neglected tropical disease. *ChemMedChem* **14**, 386-453 (2019). https://doi.org:10.1002/cmdc.201800755

47 Meulenbeld, H. J., Mathijssen, R. H., Verweij, J., de Wit, R. & de Jonge, M. J. Danusertib, an aurora kinase inhibitor. *Expert Opin Investig Drugs* **21**, 383-393 (2012). https://doi.org:10.1517/13543784.2012.652303

48 Gamo, F. J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J. L., Vanderwall, D. E., Green, D. V., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R. & Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature* **465**, 305-310 (2010). https://doi.org:10.1038/nature09107

49 Guiguemde, W. A., Shelat, A. A., Bouck, D., Duffy, S., Crowther, G. J., Davis, P. H., Smithson, D. C., Connelly, M., Clark, J., Zhu, F., Jimenez-Diaz, M. B., Martinez, M. S., Wilson, E. B., Tripathi, A. K., Gut, J., Sharlow, E. R., Bathurst, I., El Mazouni, F., Fowble, J. W., Forquer, I., McGinley, P. L., Castro, S., Angulo-Barturen, I., Ferrer, S., Rosenthal, P. J., Derisi, J. L., Sullivan, D. J., Lazo, J. S., Roos, D. S., Riscoe, M. K., Phillips, M. A., Rathod, P. K., Van Voorhis, W. C., Avery, V. M. & Guy, R. K. Chemical genetics of *Plasmodium falciparum*. *Nature* **465**, 311-315 (2010). https://doi.org:10.1038/nature09099

50 Gagaring, K., Borboa, R., Francek, C., Chen, Z., Buenviaje, J., Plouffe, D., Winzeler, E., Brinker, A., Diagana, T., Taylor, J., Glynne, R., Chatterjee, A. & Kuhen, K. Novartis-GNF Malaria Box. *Genomics Institute of the Novartis Research Foundation (GNF) and Novartis Institute for Tropical Disease* (2010).

51 Berg, M., Van der Veken, P., Joossens, J., Muthusamy, V., Breugelmans, M., Moss, C. X., Rudolf, J., Cos, P., Coombs, G. H., Maes, L., Haemers, A., Mottram, J. C. & Augustyns, K. Design and evaluation of *Trypanosoma brucei* metacaspase inhibitors. *Bioorg Med Chem Lett* **20**, 2001-2006 (2010). https://doi.org:10.1016/j.bmcl.2010.01.099

52 Murphy, R. C., Ojo, K. K., Larson, E. T., Castellanos-Gonzalez, A., Perera, B. G., Keyloun, K. R., Kim, J. E., Bhandari, J. G., Muller, N. R., Verlinde, C. L., White, A. C., Jr., Merritt, E. A., Van Voorhis, W. C. & Maly, D. J. Discovery of potent and selective inhibitors of calcium-dependent protein kinase 1 (CDPK1) from *C. parvum* and *T. gondii*. *ACS Med Chem Lett* **1**, 331-335 (2010). https://doi.org:10.1021/ml100096t

53 Albanese, C., Alzani, R., Amboldi, N., Avanzi, N., Ballinari, D., Brasca, M. G., Festuccia, C., Fiorentini, F., Locatelli, G., Pastori, W., Patton, V., Roletto, F., Colotta, F., Galvani, A., Isacchi, A., Moll, J., Pesenti, E., Mercurio, C. & Ciomei, M. Dual targeting of CDK and tropomyosin receptor kinase families by the oral inhibitor PHA-848125, an agent with broad-spectrum antitumor efficacy. *Mol Cancer Ther* **9**, 2243-2254 (2010). https://doi.org:10.1158/1535-7163.MCT-10-0190

54 Weiss, G. J., Hidalgo, M., Borad, M. J., Laheru, D., Tibes, R., Ramanathan, R. K., Blaydorn, L., Jameson, G., Jimeno, A., Isaacs, J. D., Scaburri, A., Pacciarini, M. A., Fiorentini, F., Ciomei, M. & Von Hoff, D. D. Phase I study of the safety, tolerability and pharmacokinetics of PHA-848125AC, a dual tropomyosin receptor kinase A and cyclin-dependent kinase inhibitor, in patients with advanced solid malignancies. *Invest New Drugs* **30**, 2334-2343 (2012). https://doi.org:10.1007/s10637-011-9774-6

55 Summers, R. L., Pasaje, C. F. A., Pisco, J. P., Striepen, J., Luth, M. R., Kumpornsin, K., Carpenter, E. F., Munro, J. T., Lin, D., Plater, A., Punekar, A. S., Shepherd, A. M., Shepherd, S. M., Vanaerschot, M., Murithi, J. M., Rubiano, K., Akidil, A., Ottilie, S., Mittal, N., Dilmore, A. H., Won, M., Mandt, R. E. K., McGowen, K., Owen, E., Walpole, C., Llinas, M., Lee, M. C. S., Winzeler, E. A., Fidock, D. A., Gilbert, I. H., Wirth, D. F., Niles, J. C., Baragana, B. & Lukens, A. K. Chemogenomics identifies acetyl-coenzyme A synthetase as a target for malaria treatment and prevention. *Cell Chem Biol* **29**, 191-201 e198 (2022). https://doi.org:10.1016/j.chembiol.2021.07.010

56 Crowther, G. J., Hillesland, H. K., Keyloun, K. R., Reid, M. C., Lafuente-Monasterio, M. J., Ghidelli-Disse, S., Leonard, S. E., He, P., Jones, J. C., Krahn, M. M., Mo, J. S., Dasari, K. S., Fox, A. M., Boesche, M., El Bakkouri, M., Rivas, K. L., Leroy, D., Hui, R., Drewes, G., Maly, D. J., Van Voorhis, W. C. & Ojo, K. K. Biochemical screening of five protein kinases from *Plasmodium falciparum* against 14,000 cell-active compounds. *PLoS One* **11**, e0149996 (2016). https://doi.org:10.1371/journal.pone.0149996

57 Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **47**, 2977-2980 (2004). https://doi.org:10.1021/jm030580l

58 Rout, S. & Mahapatra, R. K. In silico analysis of *Plasmodium falciparum* CDPK5 protein through molecular modeling, docking and dynamics. *J Theor Biol* **461**, 254-267 (2019). https://doi.org:10.1016/j.jtbi.2018.10.045

59 Hoepfner, D., McNamara, C. W., Lim, C. S., Studer, C., Riedl, R., Aust, T., McCormack, S. L., Plouffe, D. M., Meister, S., Schuierer, S., Plikat, U., Hartmann, N., Staedtler, F., Cotesta, S., Schmitt, E. K., Petersen, F., Supek, F., Glynne, R. J., Tallarico, J. A., Porter, J. A., Fishman, M. C., Bodenreider, C., Diagana, T. T., Movva, N. R. & Winzeler, E. A. Selective and specific inhibition of the *Plasmodium falciparum* lysyl-tRNA synthetase by the fungal secondary metabolite cladosporin. *Cell Host Microbe* **11**, 654-663 (2012). https://doi.org:10.1016/j.chom.2012.04.015

60 MalariaGen, Abdel Hamid, M. M., Abdelraheem, M. H., Acheampong, D. O., Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C., Amenga-Etego, L., Andagalu, B., Anderson, T., Andrianaranjaka, V., Aniebo, I., Aninagyei, E., Ansah, F., Ansah, P. O., Apinjoh, T., Arnaldo, P., Ashley, E., Auburn, S., Awandare, G. A., Ba, H., Baraka, V., Barry, A., Bejon, P., Bertin, G. I., Boni, M. F., Borrmann, S., Bousema, T., Bouyou-Akotet, M., Branch, O., Bull, P. C., Cheah, H., Chindavongsa, K., Chookajorn, T., Chotivanich, K., Claessens, A., Conway, D. J., Corredor, V., Courtier, E., Craig, A., D'Alessandro, U., Dama, S., Day, N., Denis, B., Dhorda, M., Diakite, M., Djimde, A., Dolecek, C., Dondorp, A., Doumbia, S., Drakeley, C., Drury, E., Duffy, P., Echeverry, D. F., Egwang, T. G., Enosse, S. M. M., Erko, B., Fairhurst, R. M., Faiz, A., Fanello, C. A., Fleharty, M., Forbes, M., Fukuda, M., Gamboa, D., Ghansah, A., Golassa, L., Goncalves, S., Harrison, G. L. A., Healy, S. A., Hendry, J. A., Hernandez-Koutoucheva, A., Hien, T. T., Hill, C. A., Hombhanje, F., Hott, A., Htut, Y., Hussein, M., Imwong, M., Ishengoma, D., Jackson, S. A., Jacob, C. G., Jeans, J., Johnson, K. J., Kamaliddin, C., Kamau, E., Keatley, J., Kochakarn, T., Konate, D. S., Konate, A., Kone, A., Kwiatkowski, D. P., Kyaw, M. P., Kyle, D., Lawniczak, M., Lee, S. K., Lemnge, M., Lim, P., Lon, C., Loua, K. M., Mandara, C. I., Marfurt, J., Marsh, K., Maude, R. J., Mayxay, M., Maiga-Ascofare, O., Miotto, O., Mita, T., Mobegi, V., Mohamed, A. O., Mokuolu, O. A., Montgomery, J., Morang'a, C. M., Mueller, I., Murie, K., Newton, P. N., Ngo Duc, T., Nguyen, T., Nguyen, T. N., Nguyen Thi Kim, T., Nguyen Van, H., Noedl, H., Nosten, F., Noviyanti, R., Ntui, V. N., Nzila, A., Ochola-Oyier, L. I., Ocholla, H., Oduro, A., Omedo, I., Onyamboko, M. A., Ouedraogo, J. B., Oyebola, K., Oyibo, W. A., Pearson, R., Peshu, N., Phyo, A. P., Plowe, C. V., Price, R. N., Pukrittayakamee, S., Quang, H. H., Randrianarivelojosia, M., Rayner, J. C., Ringwald, P., Rosanas-Urgell, A., Rovira-Vallbona, E., Ruano-Rubio, V., Ruiz, L., Saunders, D., Shayo, A., Siba, P., Simpson, V. J., Sissoko, M. S., Smith, C., Su, X. Z., Sutherland, C., Takala-Harrison, S., Talman, A., Tavul, L., Thanh, N. V., Thathy, V., Thu, A. M., Toure, M., Tshefu, A., Verra, F., Vinetz, J., Wellems, T. E., Wendler, J., White, N. J., Whitton, G., Yavo, W. & van der Pluijm, R. W. Pf7: an open dataset of *Plasmodium falciparum* genome variation in 20,000 worldwide samples. *Wellcome Open Res* **8**, 22 (2023). https://doi.org:10.12688/wellcomeopenres.18681.1

61 Rocamora, F. & Winzeler, E. A. Genomic approaches to drug resistance in malaria. *Annu Rev Microbiol* **74**, 761-786 (2020). https://doi.org:10.1146/annurev-micro-012220-064343

62 Cowen, L. E., Sanglard, D., Calabrese, D., Sirjusingh, C., Anderson, J. B. & Kohn, L. M. Evolution of drug resistance in experimental populations of *Candida albicans*. *J Bacteriol* **182**, 1515-1522 (2000). https://doi.org:10.1128/JB.182.6.1515-1522.2000

63 Smith, T. M., Youngblom, M. A., Kernien, J. F., Mohamed, M. A., Fry, S. S., Bohr, L. L., Mortimer, T. D., O'Neill, M. B. & Pepperell, C. S. Rapid adaptation of a complex trait during experimental evolution of *Mycobacterium tuberculosis*. *Elife* **11** (2022). https://doi.org:10.7554/eLife.78454

64 Ottilie, S., Luth, M. R., Hellemann, E., Goldgof, G. M., Vigil, E., Kumar, P., Cheung, A. L., Song, M., Godinez-Macias, K. P., Carolino, K., Yang, J., Lopez, G., Abraham, M., Tarsio, M., LeBlanc, E., Whitesell, L., Schenken, J., Gunawan, F., Patel, R., Smith, J., Love, M. S., Williams, R. M., McNamara, C. W., Gerwick, W. H., Ideker, T., Suzuki, Y., Wirth, D. F., Lukens, A. K., Kane, P. M., Cowen, L. E., Durrant, J. D. & Winzeler, E. A. Adaptive laboratory evolution in *S. cerevisiae* highlights role of transcription factors in fungal xenobiotic resistance. *Commun Biol* **5**, 128 (2022). https://doi.org:10.1038/s42003-022-03076-7

65 Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E., O'Brien, J., Ruano Rubio, V., MacInnis, B., Mwangi, J., Samarakoon, U., Ranford-Cartwright, L., Ferdig, M., Hayton, K., Su, X. Z., Wellems, T., Rayner, J., McVean, G. & Kwiatkowski, D. Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res* **26**, 1288-1299 (2016). https://doi.org:10.1101/gr.203711.115

66 Yang, T., Ottilie, S., Istvan, E. S., Godinez-Macias, K. P., Lukens, A. K., Baragana, B., Campo, B., Walpole, C., Niles, J. C., Chibale, K., Dechering, K. J., Llinas, M., Lee, M. C. S., Kato, N., Wyllie, S., McNamara, C. W., Gamo, F. J., Burrows, J., Fidock, D. A., Goldberg, D. E., Gilbert, I. H., Wirth, D. F., Winzeler, E. A. & Malaria Drug Accelerator, C. MalDA, accelerating malaria drug discovery. *Trends Parasitol* **37**, 493-507 (2021). https://doi.org:10.1016/j.pt.2021.01.009

67 Siqueira-Neto, J. L., Wicht, K. J., Chibale, K., Burrows, J. N., Fidock, D. A. & Winzeler, E. A. Antimalarial drug discovery: progress and approaches. *Nat Rev Drug Discov* **22**, 807-826 (2023). https://doi.org:10.1038/s41573-023-00772-9

68 Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S. & DePristo, M. A. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-11 10 33 (2013). https://doi.org:10.1002/0471250953.bi1110s43

69 Cowell, A. N., Istvan, E. S., Lukens, A. K., Gomez-Lorenzo, M. G., Vanaerschot, M., Sakata-Kato, T., Flannery, E. L., Magistrado, P., Owen, E., Abraham, M., LaMonte, G., Painter, H. J., Williams, R. M., Franco, V., Linares, M., Arriaga, I., Bopp, S., Corey, V. C., Gnadig, N. F., Coburn-Flynn, O., Reimer, C., Gupta, P., Murithi, J. M., Moura, P. A., Fuchs, O., Sasaki, E., Kim, S. W., Teng, C. H., Wang, L. T., Akidil, A., Adjalley, S., Willis, P. A., Siegel, D., Tanaseichuk, O., Zhong, Y., Zhou, Y., Llinas, M., Ottilie, S., Gamo, F. J., Lee, M. C. S., Goldberg, D. E., Fidock, D. A., Wirth, D. F. & Winzeler, E. A. Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics. *Science* **359**, 191-199 (2018). https://doi.org:10.1126/science.aan4472

70 Rocamora, F., Zhu, L., Liong, K. Y., Dondorp, A., Miotto, O., Mok, S. & Bozdech, Z. Oxidative stress and protein damage responses mediate artemisinin resistance in malaria parasites. *PLoS Pathog* **14**, e1006930 (2018). https://doi.org:10.1371/journal.ppat.1006930

71 Kidgell, C., Volkman, S. K., Daily, J., Borevitz, J. O., Plouffe, D., Zhou, Y., Johnson, J. R., Le Roch, K., Sarr, O., Ndir, O., Mboup, S., Batalov, S., Wirth, D. F. & Winzeler, E. A. A systematic map of genetic variation in *Plasmodium falciparum*. *PLoS Pathog* **2**, e57 (2006). https://doi.org:10.1371/journal.ppat.0020057

72 Huckaby, A. C., Granum, C. S., Carey, M. A., Szlachta, K., Al-Barghouthi, B., Wang, Y. H. & Guler, J. L. Complex DNA structures trigger copy number variation across the *Plasmodium falciparum* genome. *Nucleic Acids Res* **47**, 1615-1627 (2019). https://doi.org:10.1093/nar/gky1268

73 Shafik, S. H., Richards, S. N., Corry, B. & Martin, R. E. Mechanistic basis for multidrug resistance and collateral drug sensitivity conferred to the malaria parasite by polymorphisms in *Pf*MDR1 and *Pf*CRT. *PLoS Biol* **20**, e3001616 (2022). https://doi.org:10.1371/journal.pbio.3001616

74 LaMonte, G., Lim, M. Y., Wree, M., Reimer, C., Nachon, M., Corey, V., Gedeck, P., Plouffe, D., Du, A., Figueroa, N., Yeung, B., Bifani, P. & Winzeler, E. A. Mutations in the

Plasmodium falciparum cyclic amine resistance locus (*Pf*CARL) confer multidrug resistance. *mBio* **7** (2016). https://doi.org:10.1128/mBio.00696-16

75 Rottmann, M., McNamara, C., Yeung, B. K., Lee, M. C., Zou, B., Russell, B., Seitz, P., Plouffe, D. M., Dharia, N. V., Tan, J., Cohen, S. B., Spencer, K. R., Gonzalez-Paez, G. E., Lakshminarayana, S. B., Goh, A., Suwanarusk, R., Jegla, T., Schmitt, E. K., Beck, H. P., Brun, R., Nosten, F., Renia, L., Dartois, V., Keller, T. H., Fidock, D. A., Winzeler, E. A. & Diagana, T. T. Spiroindolones, a potent compound class for the treatment of malaria. *Science* **329**, 1175-1180 (2010). https://doi.org:10.1126/science.1193225

76 McNamara, C. W., Lee, M. C., Lim, C. S., Lim, S. H., Roland, J., Simon, O., Yeung, B. K., Chatterjee, A. K., McCormack, S. L., Manary, M. J., Zeeman, A. M., Dechering, K. J., Kumar, T. S., Henrich, P. P., Gagaring, K., Ibanez, M., Kato, N., Kuhen, K. L., Fischli, C., Nagle, A., Rottmann, M., Plouffe, D. M., Bursulaya, B., Meister, S., Rameh, L., Trappe, J., Haasen, D., Timmerman, M., Sauerwein, R. W., Suwanarusk, R., Russell, B., Renia, L., Nosten, F., Tully, D. C., Kocken, C. H., Glynne, R. J., Bodenreider, C., Fidock, D. A., Diagana, T. T. & Winzeler, E. A. Targeting *Plasmodium* PI(4)K to eliminate malaria. *Nature* **504**, 248-253 (2013). https://doi.org:10.1038/nature12782

77 Korsinczky, M., Chen, N., Kotecka, B., Saul, A., Rieckmann, K. & Cheng, Q. Mutations in *Plasmodium falciparum* cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site. *Antimicrob Agents Chemother* **44**, 2100-2108 (2000). https://doi.org:10.1128/AAC.44.8.2100-2108.2000

78 Guler, J. L., Freeman, D. L., Ahyong, V., Patrapuvich, R., White, J., Gujjar, R., Phillips, M. A., DeRisi, J. & Rathod, P. K. Asexual populations of the human malaria parasite, *Plasmodium falciparum*, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications. *PLoS Pathog* **9**, e1003375 (2013). https://doi.org:10.1371/journal.ppat.1003375

79 Josling, G. A., Russell, T. J., Venezia, J., Orchard, L., van Biljon, R., Painter, H. J. & Llinas, M. Dissecting the role of *Pf*AP2-G in malaria gametocytogenesis. *Nat Commun* **11**, 1503 (2020). https://doi.org:10.1038/s41467-020-15026-0

80 Claessens, A., Affara, M., Assefa, S. A., Kwiatkowski, D. P. & Conway, D. J. Culture adaptation of malaria parasites selects for convergent loss-of-function mutants. *Sci Rep* **7**, 41303 (2017). https://doi.org:10.1038/srep41303

81 de Vries, L. E., Lunghi, M., Krishnan, A., Kooij, T. W. A. & Soldati-Favre, D. Pantothenate and CoA biosynthesis in Apicomplexa and their promise as antiparasitic drug targets. *PLoS Pathog* **17**, e1010124 (2021). https://doi.org:10.1371/journal.ppat.1010124

82 Jeninga, M. D., Quinn, J. E. & Petter, M. ApiAP2 transcription factors in apicomplexan parasites. *Pathogens* **8** (2019). https://doi.org:10.3390/pathogens8020047

83 MalariaGen, Ahouidi, A., Ali, M., Almagro-Garcia, J., Amambua-Ngwa, A., Amaratunga, C., Amato, R., Amenga-Etego, L., Andagalu, B., Anderson, T. J. C., Andrianaranjaka, V., Apinjoh, T., Ariani, C., Ashley, E. A., Auburn, S., Awandare, G. A., Ba, H., Baraka, V., Barry, A. E., Bejon, P., Bertin, G. I., Boni, M. F., Borrmann, S., Bousema, T., Branch, O., Bull, P. C., Busby, G. B. J., Chookajorn, T., Chotivanich, K., Claessens, A., Conway, D., Craig, A., D'Alessandro, U., Dama, S., Day, N. P., Denis, B., Diakite, M., Djimde, A., Dolecek, C., Dondorp, A. M., Drakeley, C., Drury, E., Duffy, P., Echeverry, D. F., Egwang, T. G., Erko, B., Fairhurst, R. M., Faiz, A., Fanello, C. A., Fukuda, M. M., Gamboa, D., Ghansah, A., Golassa, L., Goncalves, S., Hamilton, W. L., Harrison, G. L. A., Hart, L., Henrichs, C., Hien, T. T., Hill, C. A., Hodgson, A., Hubbart, C., Imwong, M., Ishengoma, D. S., Jackson, S. A., Jacob, C. G., Jeffery, B., Jeffreys, A. E., Johnson, K. J., Jyothi, D., Kamaliddin, C., Kamau, E., Kekre, M., Kluczynski, K., Kochakarn, T., Konate, A., Kwiatkowski, D. P., Kyaw, M. P., Lim, P., Lon, C., Loua, K. M., Maiga-Ascofare, O., Malangone, C., Manske, M., Marfurt, J., Marsh, K., Mayxay, M., Miles, A., Miotto, O., Mobegi, V., Mokuolu, O. A., Montgomery, J., Mueller, I., Newton, P. N., Nguyen, T., Nguyen, T. N., Noedl, H., Nosten, F., Noviyanti, R., Nzila, A., Ochola-Oyier, L. I., Ocholla, H., Oduro, A., Omedo, I., Onyamboko, M. A., Ouedraogo, J. B., Oyebola, K., Pearson, R. D., Peshu, N., Phyo, A. P., Plowe, C. V., Price, R. N., Pukrittayakamee, S., Randrianarivelojosia, M., Rayner, J. C., Ringwald, P., Rockett, K. A., Rowlands, K., Ruiz, L., Saunders, D., Shayo, A., Siba, P., Simpson, V. J., Stalker, J., Su, X. Z., Sutherland, C., Takala-Harrison, S., Tavul, L., Thathy, V., Tshefu, A., Verra, F., Vinetz, J., Wellems, T. E., Wendler, J., White, N. J., Wright, I., Yavo, W. & Ye, H. An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples. *Wellcome Open Res* **6**, 42 (2021). https://doi.org:10.12688/wellcomeopenres.16168.2

84 Le, H. G., Kang, J. M., Jun, H., Lee, J., Thai, T. L., Myint, M. K., Aye, K. S., Sohn, W. M., Shin, H. J., Kim, T. S. & Na, B. K. Changing pattern of the genetic diversities of *Plasmodium falciparum* merozoite surface protein-1 and merozoite surface protein-2 in Myanmar isolates. *Malar J* **18**, 241 (2019). https://doi.org:10.1186/s12936-019-2879-7

85 Kang, J. M., Le, H. G., Vo, T. C., Naw, H., Yoo, W. G., Sohn, W. M., Trinh, N. T. M., Quang, H. H. & Na, B. K. Genetic polymorphism and natural selection of apical membrane antigen-1 in *Plasmodium falciparum* isolates from vietnam. *Genes (Basel)* **12** (2021). https://doi.org:10.3390/genes12121903

86 Zhou, Y., Tao, L., Zhou, X., Zuo, Z., Gong, J., Liu, X., Zhou, Y., Liu, C., Sang, N., Liu, H., Zou, J., Gou, K., Yang, X. & Zhao, Y. DHODH and cancer: promising prospects to be explored. *Cancer Metab* **9**, 22 (2021). https://doi.org:10.1186/s40170-021-00250-z

87 Sindhe, K. M. V., Wu, W., Legac, J., Zhang, Y. K., Easom, E. E., Cooper, R. A., Plattner, J. J., Freund, Y. R., DeRisi, J. L. & Rosenthal, P. J. *Plasmodium falciparum* resistance to a lead benzoxaborole due to blocked compound activation and altered ubiquitination or sumoylation. *mBio* **11** (2020). https://doi.org:10.1128/mBio.02640-19

88 Mandt, R. E. K., Lafuente-Monasterio, M. J., Sakata-Kato, T., Luth, M. R., Segura, D., Pablos-Tanarro, A., Viera, S., Magan, N., Ottilie, S., Winzeler, E. A., Lukens, A. K., Gamo, F. J. & Wirth, D. F. In vitro selection predicts malaria parasite resistance to dihydroorotate dehydrogenase inhibitors in a mouse infection model. *Sci Transl Med* **11** (2019). https://doi.org:10.1126/scitranslmed.aav1636

89 Meister, S., Plouffe, D. M., Kuhen, K. L., Bonamy, G. M., Wu, T., Barnes, S. W., Bopp, S. E., Borboa, R., Bright, A. T., Che, J., Cohen, S., Dharia, N. V., Gagaring, K., Gettayacamin, M., Gordon, P., Groessl, T., Kato, N., Lee, M. C., McNamara, C. W., Fidock, D. A., Nagle, A., Nam, T. G., Richmond, W., Roland, J., Rottmann, M., Zhou, B., Froissard, P., Glynne, R. J., Mazier, D., Sattabongkot, J., Schultz, P. G., Tuntland, T., Walker, J. R., Zhou, Y., Chatterjee, A., Diagana, T. T. & Winzeler, E. A. Imaging of *Plasmodium* liver stages to drive next-generation antimalarial drug discovery. *Science* **334**, 1372-1377 (2011). https://doi.org:10.1126/science.1211936

90 Herman, J. D., Rice, D. P., Ribacke, U., Silterra, J., Deik, A. A., Moss, E. L., Broadbent, K. M., Neafsey, D. E., Desai, M. M., Clish, C. B., Mazitschek, R. & Wirth, D. F. A genomic and evolutionary approach reveals non-genetic drug resistance in malaria. *Genome Biol* **15**, 511 (2014). https://doi.org:10.1186/PREACCEPT-1067113631444973

91 Xie, S. C., Gillett, D. L., Spillman, N. J., Tsu, C., Luth, M. R., Ottilie, S., Duffy, S., Gould, A. E., Hales, P., Seager, B. A., Charron, C. L., Bruzzese, F., Yang, X., Zhao, X., Huang, S. C., Hutton, C. A., Burrows, J. N., Winzeler, E. A., Avery, V. M., Dick, L. R. & Tilley, L. Target validation and identification of novel boronate inhibitors of the *Plasmodium falciparum* proteasome. *J Med Chem* **61**, 10053-10066 (2018). https://doi.org:10.1021/acs.jmedchem.8b01161

92 Schalkwijk, J., Allman, E. L., Jansen, P. A. M., de Vries, L. E., Verhoef, J. M. J., Jackowski, S., Botman, P. N. M., Beuckens-Schortinghuis, C. A., Koolen, K. M. J., Bolscher, J. M., Vos, M. W., Miller, K., Reeves, S. A., Pett, H., Trevitt, G., Wittlin, S.,

Scheurer, C., Sax, S., Fischli, C., Angulo-Barturen, I., Jimenez-Diaz, M. B., Josling, G., Kooij, T. W. A., Bonnert, R., Campo, B., Blaauw, R. H., Rutjes, F., Sauerwein, R. W., Llinas, M., Hermkens, P. H. H. & Dechering, K. J. Antimalarial pantothenamide metabolites target acetyl-coenzyme A biosynthesis in *Plasmodium falciparum*. *Sci Transl Med* **11** (2019). https://doi.org:10.1126/scitranslmed.aas9917

93 Jimenez-Diaz, M. B., Ebert, D., Salinas, Y., Pradhan, A., Lehane, A. M., Myrand-Lapierre, M. E., O'Loughlin, K. G., Shackleford, D. M., Justino de Almeida, M., Carrillo, A. K., Clark, J. A., Dennis, A. S., Diep, J., Deng, X., Duffy, S., Endsley, A. N., Fedewa, G., Guiguemde, W. A., Gomez, M. G., Holbrook, G., Horst, J., Kim, C. C., Liu, J., Lee, M. C., Matheny, A., Martinez, M. S., Miller, G., Rodriguez-Alejandre, A., Sanz, L., Sigal, M., Spillman, N. J., Stein, P. D., Wang, Z., Zhu, F., Waterson, D., Knapp, S., Shelat, A., Avery, V. M., Fidock, D. A., Gamo, F. J., Charman, S. A., Mirsalis, J. C., Ma, H., Ferrer, S., Kirk, K., Angulo-Barturen, I., Kyle, D. E., DeRisi, J. L., Floyd, D. M. & Guy, R. K. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of *Plasmodium*. *Proc Natl Acad Sci U S A* **111**, E5455-5462 (2014). https://doi.org:10.1073/pnas.1414221111

94 Baragana, B., Hallyburton, I., Lee, M. C., Norcross, N. R., Grimaldi, R., Otto, T. D., Proto, W. R., Blagborough, A. M., Meister, S., Wirjanata, G., Ruecker, A., Upton, L. M., Abraham, T. S., Almeida, M. J., Pradhan, A., Porzelle, A., Luksch, T., Martinez, M. S., Luksch, T., Bolscher, J. M., Woodland, A., Norval, S., Zuccotto, F., Thomas, J., Simeons, F., Stojanovski, L., Osuna-Cabello, M., Brock, P. M., Churcher, T. S., Sala, K. A., Zakutansky, S. E., Jimenez-Diaz, M. B., Sanz, L. M., Riley, J., Basak, R., Campbell, M., Avery, V. M., Sauerwein, R. W., Dechering, K. J., Noviyanti, R., Campo, B., Frearson, J. A., Angulo-Barturen, I., Ferrer-Bazaga, S., Gamo, F. J., Wyatt, P. G., Leroy, D., Siegl, P., Delves, M. J., Kyle, D. E., Wittlin, S., Marfurt, J., Price, R. N., Sinden, R. E., Winzeler, E. A., Charman, S. A., Bebrevska, L., Gray, D. W., Campbell, S., Fairlamb, A. H., Willis, P. A., Rayner, J. C., Fidock, D. A., Read, K. D. & Gilbert, I. H. A novel multiple-stage antimalarial agent that inhibits protein synthesis. *Nature* **522**, 315-320 (2015). https://doi.org:10.1038/nature14451

95 Tahlan, K., Wilson, R., Kastrinsky, D. B., Arora, K., Nair, V., Fischer, E., Barnes, S. W., Walker, J. R., Alland, D., Barry, C. E., 3rd & Boshoff, H. I. SQ109 targets MmpL3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of *Mycobacterium tuberculosis*. *Antimicrob Agents Chemother* **56**, 1797-1809 (2012). https://doi.org:10.1128/AAC.05708-11

96 Stokes, B. H., Yoo, E., Murithi, J. M., Luth, M. R., Afanasyev, P., da Fonseca, P. C. A., Winzeler, E. A., Ng, C. L., Bogyo, M. & Fidock, D. A. Covalent *Plasmodium falciparum*-selective proteasome inhibitors exhibit a low propensity for generating resistance in vitro

and synergize with multiple antimalarial agents. *PLoS Pathog* **15**, e1007722 (2019). https://doi.org:10.1371/journal.ppat.1007722

97 Fisher, G. M., Cobbold, S. A., Jezewski, A., Carpenter, E. F., Arnold, M., Cowell, A. N., Tjhin, E. T., Saliba, K. J., Skinner-Adams, T. S., Lee, M. C. S., Odom John, A., Winzeler, E. A., McConville, M. J., Poulsen, S. A. & Andrews, K. T. The key glycolytic enzyme phosphofructokinase is involved in resistance to Antiplasmodial glycosides. *mBio* **11** (2020). https://doi.org:10.1128/mBio.02842-20

98 Ellis, K. M., Lucantoni, L., Chavchich, M., Abraham, M., De Paoli, A., Luth, M. R., Zeeman, A. M., Delves, M. J., Teran, F. S., Straschil, U., Baum, J., Kocken, C. H. M., Ralph, S. A., Winzeler, E. A., Avery, V. M., Edstein, M. D., Baell, J. B. & Creek, D. J. The novel bis-1,2,4-triazine MIPS-0004373 Demonstrates Rapid and Potent Activity against all blood stages of the malaria parasite. *Antimicrob Agents Chemother* **65**, e0031121 (2021). https://doi.org:10.1128/AAC.00311-21

99 Gisselberg, J. E., Herrera, Z., Orchard, L. M., Llinas, M. & Yeh, E. Specific inhibition of the bifunctional farnesyl/geranylgeranyl diphosphate synthase in malaria parasites via a new small-molecule binding site. *Cell Chem Biol* **25**, 185-193 e185 (2018). https://doi.org:10.1016/j.chembiol.2017.11.010

100 Vanaerschot, M., Murithi, J. M., Pasaje, C. F. A., Ghidelli-Disse, S., Dwomoh, L., Bird, M., Spottiswoode, N., Mittal, N., Arendse, L. B., Owen, E. S., Wicht, K. J., Siciliano, G., Bosche, M., Yeo, T., Kumar, T. R. S., Mok, S., Carpenter, E. F., Giddins, M. J., Sanz, O., Ottilie, S., Alano, P., Chibale, K., Llinas, M., Uhlemann, A. C., Delves, M., Tobin, A. B., Doerig, C., Winzeler, E. A., Lee, M. C. S., Niles, J. C. & Fidock, D. A. Inhibition of resistance-refractory *P. falciparum* kinase PKG delivers prophylactic, blood stage, and transmission-blocking antiplasmodial activity. *Cell Chem Biol* **27**, 806-816 e808 (2020). https://doi.org:10.1016/j.chembiol.2020.04.001

101 Knaab, T. C., Held, J., Burckhardt, B. B., Rubiano, K., Okombo, J., Yeo, T., Mok, S., Uhlemann, A. C., Lungerich, B., Fischli, C., Pessanha de Carvalho, L., Mordmuller, B., Wittlin, S., Fidock, D. A. & Kurz, T. 3-Hydroxy-propanamidines, a new class of orally active antimalarials targeting *Plasmodium falciparum*. *J Med Chem* **64**, 3035-3047 (2021). https://doi.org:10.1021/acs.jmedchem.0c01744

102 Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., Ursos, L. M., Sidhu, A. B., Naude, B., Deitsch, K. W., Su, X. Z., Wootton, J. C., Roepe, P. D. & Wellems, T. E. Mutations in the *P. falciparum* digestive vacuole transmembrane

protein *Pf*CRT and evidence for their role in chloroquine resistance. *Mol Cell* **6**, 861-871 (2000). https://doi.org:10.1016/s1097-2765(05)00077-8

103 Goodyer, I. D. & Taraschi, T. F. *Plasmodium falciparum*: a simple, rapid method for detecting parasite clones in microtiter plates. *Exp Parasitol* **86**, 158-160 (1997). https://doi.org:10.1006/expr.1997.4156

104 Plouffe, D., Brinker, A., McNamara, C., Henson, K., Kato, N., Kuhen, K., Nagle, A., Adrian, F., Matzen, J. T., Anderson, P., Nam, T. G., Gray, N. S., Chatterjee, A., Janes, J., Yan, S. F., Trager, R., Caldwell, J. S., Schultz, P. G., Zhou, Y. & Winzeler, E. A. In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc Natl Acad Sci* **105**, 9059-9064 (2008). https://doi.org:10.1073/pnas.0802982105

105 Vanaerschot, M., Lucantoni, L., Li, T., Combrinck, J. M., Ruecker, A., Kumar, T. R. S., Rubiano, K., Ferreira, P. E., Siciliano, G., Gulati, S., Henrich, P. P., Ng, C. L., Murithi, J. M., Corey, V. C., Duffy, S., Lieberman, O. J., Veiga, M. I., Sinden, R. E., Alano, P., Delves, M. J., Lee Sim, K., Winzeler, E. A., Egan, T. J., Hoffman, S. L., Avery, V. M. & Fidock, D. A. Hexahydroquinolines are antimalarial candidates with potent blood-stage and transmission-blocking activity. *Nat Microbiol* **2**, 1403-1414 (2017). https://doi.org:10.1038/s41564-017-0007-4

106 Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012). https://doi.org:10.4161/fly.19695

107 Metwally, N. G., Tilly, A. K., Lubiana, P., Roth, L. K., Dorpinghaus, M., Lorenzen, S., Schuldt, K., Witt, S., Bachmann, A., Tidow, H., Gutsmann, T., Burmester, T., Roeder, T., Tannich, E. & Bruchhaus, I. Characterisation of *Plasmodium falciparum* populations selected on the human endothelial receptors P-selectin, E-selectin, CD9 and CD151. *Sci Rep* **7**, 4069 (2017). https://doi.org:10.1038/s41598-017-04241-3

108 Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V. & Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012). https://doi.org:10.1093/bioinformatics/bts378

109 Spillman, N. J. & Kirk, K. The malaria parasite cation ATPase *Pf*ATP4 and its role in the mechanism of action of a new arsenal of antimalarial drugs. *Int J Parasitol Drugs Drug Resist* **5**, 149-162 (2015). https://doi.org:10.1016/j.ijpddr.2015.07.001

110 Keller, T. L., Zocco, D., Sundrud, M. S., Hendrick, M., Edenius, M., Yum, J., Kim, Y. J., Lee, H. K., Cortese, J. F., Wirth, D. F., Dignam, J. D., Rao, A., Yeo, C. Y., Mazitschek, R. & Whitman, M. Halofuginone and other febrifugine derivatives inhibit prolyl-tRNA synthetase. *Nat Chem Biol* **8**, 311-317 (2012). https://doi.org:10.1038/nchembio.790

111 Kato, N., Comer, E., Sakata-Kato, T., Sharma, A., Sharma, M., Maetani, M., Bastien, J., Brancucci, N. M., Bittker, J. A., Corey, V., Clarke, D., Derbyshire, E. R., Dornan, G. L., Duffy, S., Eckley, S., Itoe, M. A., Koolen, K. M., Lewis, T. A., Lui, P. S., Lukens, A. K., Lund, E., March, S., Meibalan, E., Meier, B. C., McPhail, J. A., Mitasev, B., Moss, E. L., Sayes, M., Van Gessel, Y., Wawer, M. J., Yoshinaga, T., Zeeman, A. M., Avery, V. M., Bhatia, S. N., Burke, J. E., Catteruccia, F., Clardy, J. C., Clemons, P. A., Dechering, K. J., Duvall, J. R., Foley, M. A., Gusovsky, F., Kocken, C. H., Marti, M., Morningstar, M. L., Munoz, B., Neafsey, D. E., Sharma, A., Winzeler, E. A., Wirth, D. F., Scherer, C. A. & Schreiber, S. L. Diversity-oriented synthesis yields novel multistage antimalarial inhibitors. *Nature* **538**, 344-349 (2016). https://doi.org:10.1038/nature19804

112 Xie, S. C., Metcalfe, R. D., Dunn, E., Morton, C. J., Huang, S. C., Puhalovich, T., Du, Y., Wittlin, S., Nie, S., Luth, M. R., Ma, L., Kim, M. S., Pasaje, C. F. A., Kumpornsin, K., Giannangelo, C., Houghton, F. J., Churchyard, A., Famodimu, M. T., Barry, D. C., Gillett, D. L., Dey, S., Kosasih, C. C., Newman, W., Niles, J. C., Lee, M. C. S., Baum, J., Ottilie, S., Winzeler, E. A., Creek, D. J., Williamson, N., Parker, M. W., Brand, S., Langston, S. P., Dick, L. R., Griffin, M. D. W., Gould, A. E. & Tilley, L. Reaction hijacking of tyrosine tRNA synthetase as a new whole-of-life-cycle antimalarial strategy. *Science* **376**, 1074-1079 (2022). https://doi.org:10.1126/science.abn0611

113 Istvan, E. S., Guerra, F., Abraham, M., Huang, K. S., Rocamora, F., Zhao, H., Xu, L., Pasaje, C., Kumpornsin, K., Luth, M. R., Cui, H., Yang, T., Palomo Diaz, S., Gomez-Lorenzo, M. G., Qahash, T., Mittal, N., Ottilie, S., Niles, J., Lee, M. C. S., Llinas, M., Kato, N., Okombo, J., Fidock, D. A., Schimmel, P., Gamo, F. J., Goldberg, D. E. & Winzeler, E. A. Cytoplasmic isoleucyl tRNA synthetase as an attractive multistage antimalarial drug target. *Sci Transl Med* **15**, eadc9249 (2023). https://doi.org:10.1126/scitranslmed.adc9249

114 Hasan, S., Daugelat, S., Rao, P. S. & Schreiber, M. Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. *PLoS Comput Biol* **2**, e61 (2006). https://doi.org:10.1371/journal.pcbi.0020061

115 Magarinos, M. P., Carmona, S. J., Crowther, G. J., Ralph, S. A., Roos, D. S., Shanmugam, D., Van Voorhis, W. C. & Aguero, F. TDR Targets: a chemogenomics resource for neglected diseases. *Nucleic Acids Res* **40**, D1118-1127 (2012). https://doi.org:10.1093/nar/gkr1053

116 Ali, F., Wali, H., Jan, S., Zia, A., Aslam, M., Ahmad, I., Afridi, S. G., Shams, S. & Khan, A. Analysing the essential proteins set of *Plasmodium falciparum* PF3D7 for novel drug targets identification against malaria. *Malar J* **20**, 335 (2021). https://doi.org:10.1186/s12936-021-03865-1

117 Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. & Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303 (2018). https://doi.org:10.1093/nar/gky427

118 Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **31**, 455-461 (2010). https://doi.org:10.1002/jcc.21334

119 Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**, D439-D444 (2022). https://doi.org:10.1093/nar/gkab1061

120 Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. & Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023). https://doi.org:10.1126/science.ade2574

121 Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**, 205-213 (2023). https://doi.org:10.1038/s41592-022-01685-y

122 Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* **35**, D198-201 (2007). https://doi.org:10.1093/nar/gkl999

123 Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A. & Schomburg, D. The BRENDA enzyme information system-from a database to an expert system. *J Biotechnol* **261**, 194-206 (2017). https://doi.org:10.1016/j.jbiotec.2017.04.020

124 Zhang, M., Wang, C., Otto, T. D., Oberstaller, J., Liao, X., Adapa, S. R., Udenze, K., Bronner, I. F., Casandra, D., Mayho, M., Brown, J., Li, S., Swanson, J., Rayner, J. C., Jiang, R. H. Y. & Adams, J. H. Uncovering the essential genes of the human malaria parasite *Plasmodium falciparum* by saturation mutagenesis. *Science* **360** (2018). https://doi.org:10.1126/science.aap7847

125 Schwach, F., Bushell, E., Gomes, A. R., Anar, B., Girling, G., Herd, C., Rayner, J. C. & Billker, O. PlasmoGEM, a database supporting a community resource for large-scale experimental genetics in malaria parasites. *Nucleic Acids Res* **43**, D1176-1182 (2015). https://doi.org:10.1093/nar/gku1143

126 Janse, C. J., Kroeze, H., van Wigcheren, A., Mededovic, S., Fonager, J., Franke-Fayard, B., Waters, A. P. & Khan, S. M. A genotype and phenotype database of genetically modified malaria-parasites. *Trends Parasitol* **27**, 31-39 (2011). https://doi.org:10.1016/j.pt.2010.06.016

127 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003). https://doi.org:10.1101/gr.1224503

128 Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V. & Zdobnov, E. M. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445-D451 (2023). https://doi.org:10.1093/nar/gkac998

129 Altenhoff, A. M., Warwick Vesztrocy, A., Bernard, C., Train, C. M., Nicheperovich, A., Prieto Banos, S., Julca, I., Moi, D., Nevers, Y., Majidian, S., Dessimoz, C. & Glover, N. M. OMA orthology in 2024: improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res* **52**, D513-D521 (2024). https://doi.org:10.1093/nar/gkad1020

130 Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The protein data bank. *Nucleic Acids Res* **28**, 235-242 (2000). https://doi.org:10.1093/nar/28.1.235

131 Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Stoeckert, C. J., Jr., Treatman, C. & Wang, H. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* **37**, D539-543 (2009). https://doi.org:10.1093/nar/gkn814

132 van Beusekom, B., Touw, W. G., Tatineni, M., Somani, S., Rajagopal, G., Luo, J., Gilliland, G. L., Perrakis, A. & Joosten, R. P. Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Sci* **27**, 798-808 (2018). https://doi.org:10.1002/pro.3353

133 Penel, S., Arigon, A. M., Dufayard, J. F., Sertier, A. S., Daubin, V., Duret, L., Gouy, M. & Perriere, G. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* **10 Suppl 6**, S3 (2009). https://doi.org:10.1186/1471-2105-10-S6-S3

134 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). https://doi.org:10.1186/1471-2105-10-421

135 Akter, M., Drinkwater, N., Devine, S. M., Drew, S. C., Krishnarjuna, B., Debono, C. O., Wang, G., Scanlon, M. J., Scammells, P. J., McGowan, S., MacRaild, C. A. & Norton, R. S. Identification of the binding site of apical membrane antigen 1 (AMA1) inhibitors using a paramagnetic probe. *ChemMedChem* **14**, 603-612 (2019). https://doi.org:10.1002/cmdc.201800802

136 N'Dri, M. E., Royer, L. & Lavazec, C. Tadalafil impacts the mechanical properties of *Plasmodium falciparum* gametocyte-infected erythrocytes. *Mol Biochem Parasitol* **244**, 111392 (2021). https://doi.org:10.1016/j.molbiopara.2021.111392

137 Freitas-Junior, L. H., Bottius, E., Pirrit, L. A., Deitsch, K. W., Scheidig, C., Guinet, F., Nehrbass, U., Wellems, T. E. & Scherf, A. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407**, 1018-1022 (2000). https://doi.org:10.1038/35039531

138 Taylor, H. M., Kyes, S. A. & Newbold, C. I. Var gene diversity in *Plasmodium falciparum* is generated by frequent recombination events. *Mol Biochem Parasitol* **110**, 391-397 (2000). https://doi.org:10.1016/s0166-6851(00)00286-3

139 Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D. & Sherry, S. T. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-D26 (2022). https://doi.org:10.1093/nar/gkab1112

140 Howick, V. M., Russell, A. J. C., Andrews, T., Heaton, H., Reid, A. J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L. H., Rayner, J. C., Berriman, M., Herren, J. K., Billker, O., Hemberg, M., Talman, A. M. & Lawniczak, M. K. N. The Malaria Cell Atlas: single parasite transcriptomes across the complete *Plasmodium* life cycle. *Science* **365** (2019). https://doi.org:10.1126/science.aaw2619

141 Le Roch, K. G., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., De La Vega, P., Holder, A. A., Batalov, S., Carucci, D. J. & Winzeler, E. A. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science* **301**, 1503-1508 (2003). https://doi.org:10.1126/science.1087025

142 Milani, M., Balconi, E., Aliverti, A., Mastrangelo, E., Seeber, F., Bolognesi, M. & Zanetti, G. Ferredoxin-NADP+ reductase from *Plasmodium falciparum* undergoes NADP+-dependent dimerization and inactivation: functional and crystallographic analysis. *J Mol Biol* **367**, 501-513 (2007). https://doi.org:10.1016/j.jmb.2007.01.005

143 Wang, C., Zhang, B., Kruger, A., Du, X., Visser, L., Domling, A. S. S., Wrenger, C. & Groves, M. R. Discovery of small-molecule allosteric inhibitors of *Pf*ATC as antimalarials. *J Am Chem Soc* **144**, 19070-19077 (2022). https://doi.org:10.1021/jacs.2c08128

144 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309 (2005). https://doi.org:10.1093/nar/gki524

145 van Leeuwen, J., Pons, C., Tan, G., Wang, Z. Y., Hou, J., Weile, J., Gebbia, M., Liang, W., Shuteriqi, E., Li, Z., Lopes, M., Usaj, M., Dos Santos Lopes, A., van Lieshout, N., Myers, C. L., Roth, F. P., Aloy, P., Andrews, B. J. & Boone, C. Systematic analysis of bypass

suppression of essential genes. *Mol Syst Biol* **16**, e9828 (2020). https://doi.org:10.15252/msb.20209828

146 Hitz, E., Gruninger, O., Passecker, A., Wyss, M., Scheurer, C., Wittlin, S., Beck, H. P., Brancucci, N. M. B. & Voss, T. S. The catalytic subunit of *Plasmodium falciparum* casein kinase 2 is essential for gametocytogenesis. *Commun Biol* **4**, 336 (2021). https://doi.org:10.1038/s42003-021-01873-0

147 Thommen, B. T., Dziekan, J. M., Achcar, F., Tjia, S., Passecker, A., Buczak, K., Gumpp, C., Schmidt, A., Rottmann, M., Gruring, C., Marti, M., Bozdech, Z. & Brancucci, N. M. B. Genetic validation of *Pf*FKBP35 as an antimalarial drug target. *Elife* **12** (2023). https://doi.org:10.7554/eLife.86975

148 Valenzuela, M. V., Domenech, M., Mateos-Martinez, P., Gonzalez-Camacho, F., de la Campa, A. G. & Garcia, M. T. Antibacterial activity of a DNA topoisomerase I inhibitor versus fluoroquinolones in *Streptococcus pneumoniae*. *PLoS One* **15**, e0241780 (2020). https://doi.org:10.1371/journal.pone.0241780

149 Allen, S. M., Lim, E. E., Jortzik, E., Preuss, J., Chua, H. H., MacRae, J. I., Rahlfs, S., Haeussler, K., Downton, M. T., McConville, M. J., Becker, K. & Ralph, S. A. *Plasmodium falciparum* glucose-6-phosphate dehydrogenase 6-phosphogluconolactonase is a potential drug target. *FEBS J* **282**, 3808-3823 (2015). https://doi.org:10.1111/febs.13380

150 Pakosz, Z., Lin, T. Y., Michalczyk, E., Nagano, S. & Heddle, J. G. Inhibitory compounds targeting *Plasmodium falciparum* gyrase b. *Antimicrob Agents Chemother* **65**, e0026721 (2021). https://doi.org:10.1128/AAC.00267-21

151 Dar, A., Godara, P., Prusty, D. & Bashir, M. *Plasmodium falciparum* topoisomerases: emerging targets for anti-malarial therapy. *Eur J Med Chem* **265**, 116056 (2024). https://doi.org:10.1016/j.ejmech.2023.116056

152 Wang, C., Kruger, A., Du, X., Wrenger, C. & Groves, M. R. Novel highlight in malarial drug discovery: aspartate transcarbamoylase. *Front Cell Infect Microbiol* **12**, 841833 (2022). https://doi.org:10.3389/fcimb.2022.841833

153 Canon, J., Rex, K., Saiki, A. Y., Mohr, C., Cooke, K., Bagal, D., Gaida, K., Holt, T., Knutson, C. G., Koppada, N., Lanman, B. A., Werner, J., Rapaport, A. S., San Miguel, T., Ortiz, R., Osgood, T., Sun, J. R., Zhu, X., McCarter, J. D., Volak, L. P., Houk, B. E., Fakih,

M. G., O'Neil, B. H., Price, T. J., Falchook, G. S., Desai, J., Kuo, J., Govindan, R., Hong, D. S., Ouyang, W., Henary, H., Arvedson, T., Cee, V. J. & Lipford, J. R. The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217-223 (2019). https://doi.org:10.1038/s41586-019-1694-1

154 Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548-551 (2013). https://doi.org:10.1038/nature12796

155 Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M. & Whitty, A. Cryptic binding sites on proteins: definition, detection, and druggability. *Curr Opin Chem Biol* **44**, 1-8 (2018). https://doi.org:10.1016/j.cbpa.2018.05.003

156 Meller, A., Ward, M., Borowsky, J., Kshirsagar, M., Lotthammer, J. M., Oviedo, F., Ferres, J. L. & Bowman, G. R. Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Nat Commun* **14**, 1177 (2023). https://doi.org:10.1038/s41467-023-36699-3

157 Smith, R. D. & Carlson, H. A. Identification of cryptic binding sites using MixMD with standard and accelerated molecular dynamics. *J Chem Inf Model* **61**, 1287-1299 (2021). https://doi.org:10.1021/acs.jcim.0c01002

158 Kuzmanic, A., Bowman, G. R., Juarez-Jimenez, J., Michel, J. & Gervasio, F. L. Investigating cryptic binding sites by molecular dynamics simulations. *Acc Chem Res* **53**, 654-661 (2020). https://doi.org:10.1021/acs.accounts.9b00613

159 Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., Whitty, A. & Vajda, S. Exploring the structural origins of cryptic sites on proteins. *Proc Natl Acad Sci U S A* **115**, E3416-E3425 (2018). https://doi.org:10.1073/pnas.1711490115

160 Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., Schneidman-Duhovny, D., Demerdash, O. N., Mitchell, J. C., Wells, J. A., Fraser, J. S. & Sali, A. CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *J Mol Biol* **428**, 709-719 (2016). https://doi.org:10.1016/j.jmb.2016.01.029

161 Xie, X., Yu, T., Li, X., Zhang, N., Foster, L. J., Peng, C., Huang, W. & He, G. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal Transduct Target Ther* **8**, 335 (2023). https://doi.org:10.1038/s41392-023-01589-z

162 Sayers, E. A general introduction to the E-utilities. *Bethesda (MD): National Center for Biotechnology Information (US)*, Available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/ (2009).

163 McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr F Struct Biol Commun* **70**, 2-20 (2014). https://doi.org:10.1107/S2053230X13033141

164 Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramirez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P. & Tang, H. GOATOOLS: a python library for gene ontology analyses. *Sci Rep* **8**, 10872 (2018). https://doi.org:10.1038/s41598-018-28948-z

165 Dubach, R. A. & Dubach, J. M. Autocorrelation analysis of a phenotypic screen reveals hidden drug activity. *Sci Rep* **14**, 10046 (2024). https://doi.org:10.1038/s41598-024-60654-x

166 Challis, M. P., Devine, S. M. & Creek, D. J. Current and emerging target identification methods for novel antimalarials. *Int J Parasitol Drugs Drug Resist* **20**, 135-144 (2022). https://doi.org:10.1016/j.ijpddr.2022.11.001

167 Schmitt, E. K., Ndayisaba, G., Yeka, A., Asante, K. P., Grobusch, M. P., Karita, E., Mugerwa, H., Asiimwe, S., Oduro, A., Fofana, B., Doumbia, S., Su, G., Csermak Renner, K., Venishetty, V. K., Sayyed, S., Straimer, J., Demin, I., Barsainya, S., Boulton, C. & Gandhi, P. Efficacy of cipargamin (KAE609) in a randomized, phase II dose-escalation study in adults in Sub-Saharan Africa with uncomplicated *Plasmodium falciparum* malaria. *Clin Infect Dis* **74**, 1831-1839 (2022). https://doi.org:10.1093/cid/ciab716

168 Tabana, Y., Babu, D., Fahlman, R., Siraki, A. G. & Barakat, K. Target identification of small molecules: an overview of the current applications in drug discovery. *BMC Biotechnol* **23**, 44 (2023). https://doi.org:10.1186/s12896-023-00815-4

169 Modrzynska, K., Pfander, C., Chappell, L., Yu, L., Suarez, C., Dundas, K., Gomes, A. R., Goulding, D., Rayner, J. C., Choudhary, J. & Billker, O. A knockout screen of ApiAP2 genes reveals networks of interacting transcriptional regulators controlling the *Plasmodium*

life cycle. *Cell Host Microbe* **21**, 11-22 (2017). https://doi.org:10.1016/j.chom.2016.12.003

170 Peterson, E. J., Reiss, D. J., Turkarslan, S., Minch, K. J., Rustad, T., Plaisier, C. L., Longabaugh, W. J., Sherman, D. R. & Baliga, N. S. A high-resolution network model for global gene regulation in *Mycobacterium tuberculosis*. *Nucleic Acids Res* **42**, 11291-11303 (2014). https://doi.org:10.1093/nar/gku777

171 Wainberg, M., Kamber, R. A., Balsubramani, A., Meyers, R. M., Sinnott-Armstrong, N., Hornburg, D., Jiang, L., Chan, J., Jian, R., Gu, M., Shcherbina, A., Dubreuil, M. M., Spees, K., Meuleman, W., Snyder, M. P., Bassik, M. C. & Kundaje, A. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. *Nat Genet* **53**, 638-649 (2021). https://doi.org:10.1038/s41588-021-00840-z

172 Wacker, S. A., Houghtaling, B. R., Elemento, O. & Kapoor, T. M. Using transcriptome sequencing to identify mechanisms of drug action and resistance. *Nat Chem Biol* **8**, 235-237 (2012). https://doi.org:10.1038/nchembio.779

173 Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., Murino, L., Tagliaferri, R., Brunetti-Pierri, N., Isacchi, A. & di Bernardo, D. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci U S A* **107**, 14621-14626 (2010). https://doi.org:10.1073/pnas.1000138107

174 Mok, S., Stokes, B. H., Gnadig, N. F., Ross, L. S., Yeo, T., Amaratunga, C., Allman, E., Solyakov, L., Bottrill, A. R., Tripathi, J., Fairhurst, R. M., Llinas, M., Bozdech, Z., Tobin, A. B. & Fidock, D. A. Artemisinin-resistant K13 mutations rewire *Plasmodium falciparum*'s intra-erythrocytic metabolic program to enhance survival. *Nat Commun* **12**, 530 (2021). https://doi.org:10.1038/s41467-020-20805-w

175 Tewari, S. G., Elahi, R., Kwan, B., Rajaram, K., Bhatnagar, S., Reifman, J., Prigge, S. T., Vaidya, A. B. & Wallqvist, A. Metabolic responses in blood-stage malaria parasites associated with increased and decreased sensitivity to *Pf*ATP4 inhibitors. *Malar J* **22**, 56 (2023). https://doi.org:10.1186/s12936-023-04481-x

176 Lee, H. J., Georgiadou, A., Otto, T. D., Levin, M., Coin, L. J., Conway, D. J. & Cunnington, A. J. Transcriptomic studies of malaria: a paradigm for investigation of systemic host-pathogen interactions. *Microbiol Mol Biol Rev* **82** (2018). https://doi.org:10.1128/MMBR.00071-17

177 LaMonte, G. M., Rocamora, F., Marapana, D. S., Gnadig, N. F., Ottilie, S., Luth, M. R., Worgall, T. S., Goldgof, G. M., Mohunlal, R., Santha Kumar, T. R., Thompson, J. K., Vigil, E., Yang, J., Hutson, D., Johnson, T., Huang, J., Williams, R. M., Zou, B. Y., Cheung, A. L., Kumar, P., Egan, T. J., Lee, M. C. S., Siegel, D., Cowman, A. F., Fidock, D. A. & Winzeler, E. A. Pan-active imidazolopiperazine antimalarials target the *Plasmodium falciparum* intracellular secretory pathway. *Nat Commun* **11**, 1780 (2020). https://doi.org:10.1038/s41467-020-15440-4

178 Winzeler, E., Carolino, K., De Souza, M. L., Chen, D., Farre, J. C., Blauwkamp, J., Absalon, S., Ghidelli-Disse, S., Morano, A., Dvorin, J., Lafuente-Monasterio, M. J. & Gamo, F. J. *Plasmodium* SEY1 is a novel druggable target that contributes to imidazolopiperazine mechanism of action. *Res Sq* (2024). https://doi.org:10.21203/rs.3.rs-4892449/v1

179 Wells, T. N. Natural products as starting points for future anti-malarial therapies: going back to our roots? *Malar J* **10 Suppl 1**, S3 (2011). https://doi.org:10.1186/1475-2875-10-S1-S3

180 Maan, H., Zhang, L., Yu, C., Geuenich, M. J., Campbell, K. R. & Wang, B. Characterizing the impacts of dataset imbalance on single-cell data integration. *Nat Biotechnol* **42**, 1899-1908 (2024). https://doi.org:10.1038/s41587-023-02097-9

181 Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C. & Satija, R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293-304 (2024). https://doi.org:10.1038/s41587-023-01767-y

182 Cuevas-Diaz Duran, R., Wei, H. & Wu, J. Data normalization for addressing the challenges in the analysis of single-cell transcriptomic datasets. *BMC Genomics* **25**, 444 (2024). https://doi.org:10.1186/s12864-024-10364-5

183 Kiszewski, A. E. Blocking *Plasmodium falciparum* malaria transmission with drugs: the gametocytocidal and sporontocidal properties of current and prospective antimalarials. *Pharmaceuticals* **4**, 44-68 (2010). https://doi.org:10.3390/ph4010044

184 Bunnik, E. M., Cook, K. B., Varoquaux, N., Batugedara, G., Prudhomme, J., Cort, A., Shi, L., Andolina, C., Ross, L. S., Brady, D., Fidock, D. A., Nosten, F., Tewari, R., Sinnis, P., Ay, F., Vert, J. P., Noble, W. S. & Le Roch, K. G. Changes in genome organization of

parasite-specific gene families during the *Plasmodium* transmission stages. *Nat Commun* **9**, 1910 (2018). https://doi.org:10.1038/s41467-018-04295-5

185 Shonhai, A. The role of Hsp70s in the development and pathogenicity of *Plasmodium falciparum*. *Adv Exp Med Biol* **1340**, 75-95 (2021). https://doi.org:10.1007/978-3-030-78397-6_3

186 Spielmann, T., Fergusen, D. J. & Beck, H. P. ETRAMPs, a new *Plasmodium falciparum* gene family coding for developmentally regulated and highly charged membrane proteins located at the parasite-host cell interface. *Mol Biol Cell* **14**, 1529-1544 (2003). https://doi.org:10.1091/mbc.e02-04-0240

187 Liu, Y., Cheng, S., He, G., He, D., Wang, D., Wang, S., Chen, L., Zhu, L., Feng, Y., Cui, L., Cao, Y. & Zhu, X. An inner membrane complex protein IMC1g in *Plasmodium berghei* is involved in asexual stage schizogony and parasite transmission. *mBio* **16**, e0265224 (2025). https://doi.org:10.1128/mbio.02652-24

188 Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., Benner, C. & Chanda, S. K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019). https://doi.org:10.1038/s41467-019-09234-6

189 Studer, R. A. & Robinson-Rechavi, M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet* **25**, 210-216 (2009). https://doi.org:10.1016/j.tig.2009.03.004

190 Mishra, A., Rajput, S., Srivastava, P. N., Shabeer Ali, H. & Mishra, S. Autophagy protein Atg7 is essential for maintaining malaria parasite cellular homeostasis and organelle biogenesis. *mBio* **16**, e0273524 (2025). https://doi.org:10.1128/mbio.02735-24

191 Shcherbik, N. & Pestov, D. G. The impact of oxidative stress on ribosomes: from injury to regulation. *Cells* **8** (2019). https://doi.org:10.3390/cells8111379

192 Panchal, M., Rawat, K., Kumar, G., Kibria, K. M., Singh, S., Kalamuddin, M., Mohmmed, A., Malhotra, P. & Tuteja, R. *Plasmodium falciparum* signal recognition particle components and anti-parasitic effect of ivermectin in blocking nucleo-cytoplasmic shuttling of SRP. *Cell Death Dis* **5**, e994 (2014). https://doi.org:10.1038/cddis.2013.521

193 Andrews, S. FastQC: a quality control tool for high throughput sequence data, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).

194 Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J. & Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017). https://doi.org:10.1038/ncomms14049

195 Tadesse, F. G., Lanke, K., Nebie, I., Schildkraut, J. A., Goncalves, B. P., Tiono, A. B., Sauerwein, R., Drakeley, C., Bousema, T. & Rijpma, S. R. Molecular markers for sensitive detection of *Plasmodium falciparum* asexual stage parasites and their application in a malaria clinical trial. *Am J Trop Med Hyg* **97**, 188-198 (2017). https://doi.org:10.4269/ajtmh.16-0893

196 Nilsson, S., Angeletti, D., Wahlgren, M., Chen, Q. & Moll, K. *Plasmodium falciparum* antigen 332 is a resident peripheral membrane protein of Maurer's clefts. *PLoS One* **7**, e46980 (2012). https://doi.org:10.1371/journal.pone.0046980

197 Raj, D. K., Das Mohapatra, A., Jnawali, A., Zuromski, J., Jha, A., Cham-Kpu, G., Sherman, B., Rudlaff, R. M., Nixon, C. E., Hilton, N., Oleinikov, A. V., Chesnokov, O., Merritt, J., Pond-Tor, S., Burns, L., Jolly, G., Ben Mamoun, C., Kabyemela, E., Muehlenbachs, A., Lambert, L., Orr-Gonzalez, S., Gnadig, N. F., Fidock, D. A., Park, S., Dvorin, J. D., Pardi, N., Weissman, D., Mui, B. L., Tam, Y. K., Friedman, J. F., Fried, M., Duffy, P. E. & Kurtis, J. D. Anti-*Pf*GARP activates programmed cell death of parasites and reduces severe malaria. *Nature* **582**, 104-108 (2020). https://doi.org:10.1038/s41586-020-2220-1

198 Lopez-Barragan, M. J., Lemieux, J., Quinones, M., Williamson, K. C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K. & Su, X. Z. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics* **12**, 587 (2011). https://doi.org:10.1186/1471-2164-12-587

199 Waller, K. L., Stubberfield, L. M., Dubljevic, V., Buckingham, D. W., Mohandas, N., Coppel, R. L. & Cooke, B. M. Interaction of the exported malaria protein *Pf*332 with the red blood cell membrane skeleton. *Biochim Biophys Acta* **1798**, 861-871 (2010). https://doi.org:10.1016/j.bbamem.2010.01.018

200 Chattopadhyay, D., Rayner, J., McHenry, A. M. & Adams, J. H. The structure of the *Plasmodium falciparum* EBA175 ligand domain and the molecular basis of host specificity. *Trends Parasitol* **22**, 143-145 (2006). https://doi.org:10.1016/j.pt.2006.02.007

201 Cowman, A. F., Berry, D. & Baum, J. The cellular and molecular basis for malaria parasite invasion of the human red blood cell. *J Cell Biol* **198**, 961-971 (2012). https://doi.org:10.1083/jcb.201206112

202 Scherf, A., Carter, R., Petersen, C., Alano, P., Nelson, R., Aikawa, M., Mattei, D., Pereira da Silva, L. & Leech, J. Gene inactivation of *Pf*11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. *EMBO J* **11**, 2293-2301 (1992). https://doi.org:10.1002/j.1460-2075.1992.tb05288.x

203 Carter, R., Graves, P. M., Creasey, A., Byrne, K., Read, D., Alano, P. & Fenton, B. *Plasmodium falciparum*: an abundant stage-specific protein expressed during early gametocyte development. *Exp Parasitol* **69**, 140-149 (1989). https://doi.org:10.1016/0014-4894(89)90182-3

204 Dechering, K. J., Thompson, J., Dodemont, H. J., Eling, W. & Konings, R. N. Developmentally regulated expression of *Pf*s16, a marker for sexual differentiation of the human malaria parasite *Plasmodium falciparum*. *Mol Biochem Parasitol* **89**, 235-244 (1997). https://doi.org:10.1016/s0166-6851(97)00123-0

205 Lal, K., Delves, M. J., Bromley, E., Wastling, J. M., Tomley, F. M. & Sinden, R. E. *Plasmodium* male development gene-1 (MDV-1) is important for female sexual development and identifies a polarised plasma membrane during zygote development. *Int J Parasitol* **39**, 755-761 (2009). https://doi.org:10.1016/j.ijpara.2008.11.008