LARGE–SCALE DATA–DRIVEN NETWORK ANALYSIS OF HUMAN–*PLASMODIUM FALCIPARUM* INTERACTOME: EXTRACTING ESSENTIAL TARGETS AND PROCESSES FOR MALARIA DRUG DISCOVERY

Thesis Presented for the Degree of
MSc Med Human Genetics
by
Francis Edem Agamah
in the Division of Human Genetics
University of Cape Town

Supervised by: Prof. Emile R. Chimusa (Division of Human Genetics, UCT)
Co-supervised by: Dr. Gaston Mazandu (Division of Human of Genetics, UCT)
November 2019

Student Number: AGMFRA001

Declaration

I, the undersigned, hereby declare that this master's thesis entitled, "Large–scale data–driven network analysis of human–*Plasmodium falciparum* interactome: Extracting essential targets and processes for malaria drug discovery", submitted to the University of Cape Town, is solely my original work. I have not previously in its entirety or in part submitted it at any university for a degree. I have also acknowledged all authors' concepts and referenced direct quotations from their works.

Signature: Signed by candidate .....

Date: 25/11/2019

Abstract

Background:
Plasmodium falciparum malaria is an infectious disease considered to have great impact on public health due to its associated high mortality rates especially in sub Saharan Africa. *Falciparum* drug-resistant strains, notably, to chloroquine and sulfadoxine-pyrimethamine in Africa is traced mainly to Southeast Asia where artemisinin resistance rate is increasing. Although careful surveillance to monitor the emergence and spread of artemisinin-resistant parasite strains in Africa is on-going, research into new drugs, particularly, for African populations, is critical since there is no replaceable drug for artemisinin combination therapies (ACTs) yet.

Objective:
The overall objective of this study is to identify potential protein targets through host–pathogen protein–protein functional interaction network analysis to understand the underlying mechanisms of drug failure and identify those essential targets that can play their role in predicting potential drug candidates specific to the African populations through a protein-based approach of both host and *Plasmodium falciparum* genomic analysis.

Methods:
We leveraged malaria-specific genome wide association study summary statistics data obtained from Gambia, Kenya and Malawi populations, *Plasmodium falciparum* selective pressure variants and functional datasets (protein sequences, interologs, host-pathogen intra-organism and host-pathogen inter-organism protein-protein interactions (PPIs)) from various sources (STRING, Reactome, HPID, Uniprot, IntAct and literature) to construct overlapping functional network for both host and pathogen. Developed algorithms and a large-scale data-driven computational framework were used in this study to analyze the datasets and the constructed networks to identify densely connected subnetworks or hubs essential for network stability and integrity. The host-pathogen network was analyzed to elucidate the influence of parasite candidate key proteins within the network and predict possible resistant pathways due to host-pathogen candidate key protein interactions. We performed biological and pathway enrichment analysis on critical proteins identified to elucidate their functions. In order to leverage disease-target-drug relationships to identify potential repurposable already approved drug candidates that could be used to treat malaria, pharmaceutical datasets from drug bank were explored using semantic similarity approach based of target–associated biological processes

Results:
About 600,000 significant SNPs (p-value<0.05) from the summary statistics data were mapped to their associated genes, and we identified 79 human-associated malaria genes. The assembled parasite network comprised of 8 clusters containing 799 functional interactions between 155 reviewed proteins of which 5 clusters contained 43 key proteins (selective variants) and 2 clusters contained 2 candidate key proteins(key proteins characterized by high centrality measure), C6KTB7 and C6KTD2.
The human network comprised of 32 clusters containing 4,133,136 interactions between 20,329 unique reviewed proteins of which 7 clusters contained 760 key proteins and 2 clusters contained 6 significant human malaria-associated candidate key proteins or genes P22301 (*IL10*), P05362 (*ICAM1*), P01375 (*TNF*), P30480 (*HLA-B*), P16284 (*PECAM1*), O00206 (*TLR4*).
The generated host-pathogen network comprised of 31,512 functional interactions between 8,023 host and pathogen proteins. We also explored the association of *pfk13* gene within the host-pathogen. We observed that *pfk13* cluster with host kelch–like proteins and other regulatory genes but no direct association with our identified host candidate key malaria targets.
We implemented semantic similarity based approach complemented by Kappa and Jaccard statistical measure to identify 115 malaria–similar diseases and 26 potential repurposable drug hits that can be

appropriated experimentally for malaria treatment.

Conclusion:
In this study, we reviewed existing antimalarial drugs and resistance–associated variants contributing to the diminished sensitivity of antimalarials, especially chloroquine, sulfadoxine-pyrimethamine and artemisinin combination therapy within the African population. We also described various computational techniques implemented in predicting drug targets and leads in drug research. In our data analysis, we showed that possible mechanisms of resistance to artemisinin in Africa may arise from the combinatorial effects of many resistant genes to chloroquine and sulfadoxine–pyrimethamine. We investigated the role of *pfk13* within the host–pathogen network. We predicted key targets that have been proposed to be essential for malaria drug and vaccine development through structural and functional analysis of host and pathogen function networks. Based on our analysis, we propose these targets as essential co-targets for combinatorial malaria drug discovery.

Acknowledgment

Firstly, I would like to acknowledge Prof. Emile R. Chimusa and Dr. Gaston K. Mazandu, my supervisors, for their great support in my journey of masters' education. I thank them for their consistent tolerance, coaching and guidance in bioinformatics, advanced data analysis techniques and writing skill development.

Secondly, I would like to appreciate my mentor and supervisor, Dr. Anita Ghansah, for her good mentorship, encouragement and readiness to help in both academics and other matters. I am highly grateful to her for enduring and entertaining my continuous pressures via calls and emails.

My sincere gratitude to the HumGen family at the University of Cape Town, especially to Dr. Nicholas Ekow Thomford and members of the Chimusa's group, for their wonderful support, critiques and encouragement to study and achieve desired results.

Also, I am very grateful to Developing Excellence in Leadership Training and Malaria Elimination in sub-Saharan Africa (DELGEME) under the leadership of Prof. Abdoulaye Djimde, for this great opportunity, exposure and financial support in pursuit of this master's education.

Special thanks to the H3Africa Consortium Coordinating Center (H3ACC) under the leadership of Dr. Michelle Skelton for the support, exposure and training opportunities pertaining to genomic research during my masters.

Many thanks to the authors and developers who made available various heterogeneous datasets and tools implemented in this research. This thesis was typed out using LATEX running on an ubuntu operating computer.

I also acknowledge Center for High Performance Computing (CHPC) for their support by providing portable batch servers for running computationally intensive jobs and storing data for my research.

Finally, I am highly indebted to my family, particularly my mum, for the various support and motivation.

List of related publication

1. Francis E. Agamah, Gaston K. Mazandu, Radia Hassan, Christian D. Bope, Nicholas E. Thomford, Anita Ghansah, Emile R. Chimusa.
Computational / in silico methods in drug target and leads prediction. Briefings in Bioinformatics (2019), https://doi.org/10.1093/bib/bbz103.

2. Thomford, Nicholas Ekow, Christian Domilongo Bope, Francis Edem Agamah, Kevin Dzobo, Richmond Owusu Ateko, Emile Chimusa, Gaston Kuzamunu Mazandu, Simon Badibanga Ntumba, Collet Dandara, and Ambroise Wonkam.
Implementing Artificial Intelligence and Digital Health in Resource-Limited Settings? Top 10 Lessons We Learned in Congenital Heart Defects and Cardiology. Omics: a journal of integrative biology (2019), https://doi.org/10.1089/omi.2019.0142.

3. Francis E. Agamah, Delesa Damena, Michelle Skelton, Anita Ghansah, Gaston K. Mazandu, Emile R. Chimusa.
Network-driven analysis of human-Plasmodium falciparum interactome: processes for malaria drug discovery and extracting in silico targets. Submitted to Journal of Human Genetics.

4. Francis E. Agamah, Gaston K. Mazandu, Anita Ghansah, Emile R. Chimusa.
The emergence and spread of artemisinin resistance in Africa: Lessons from sulfadoxine-pyrimethamine and chloroquine. Under review 2019.

List of Figures

List of Tables

Contents

CHAPTER 1

1 Introduction and background

1.1 Overview

Malaria is a life-threatening mosquito-borne blood disease which is considered to have the greatest impact on public health. This parasitic disease is one of the leading worldwide infectious diseases with a devastating mortality rate especially in developing countries [1]. It is caused by *Plasmodium*, a protozoan parasite of phylum Apicomplexa.

Malaria aetiology is attributed to environmental factors, parasite virulence as well as the level of immunity by host (human) genetics [2]. The disease is primarily transmitted to humans through the bite of an infective female mosquito from the genus *Anopheles*. However, because the malaria parasite is found in erythrocytes of infected peoples, secondary transmission is through blood transfusion, organ transplant and transmission from a mother to her unborn baby prior or during delivery. The global effects and distribution (**Figure 1**) of this disease is vast and as such, threatens public health and productivity on a broad scale. It also impedes the progress of many countries especially endemic ones characterized by high poverty indices. This public health problem is aggravated by the emergence and widespread of drug-resistant parasites.

It is hypothesized that death from malaria heightened with the development of the agriculture and human habitation about 5,000 to 10,000 years ago [2]. These factors are thought to have increased human population density and also mosquitoes, the malaria vector promoting transmission. According to World Malaria Report [3], it is estimated that, 3.4 billion people in 92 countries are at risk of being infected with malaria. The burden is heaviest in the Africa, a continent where the disease brings about severe morbidity and mortality cases which significantly affects her socio-economic development. This continent is characterized by high level of genetic variation, population structure and naturally acquired immunity.

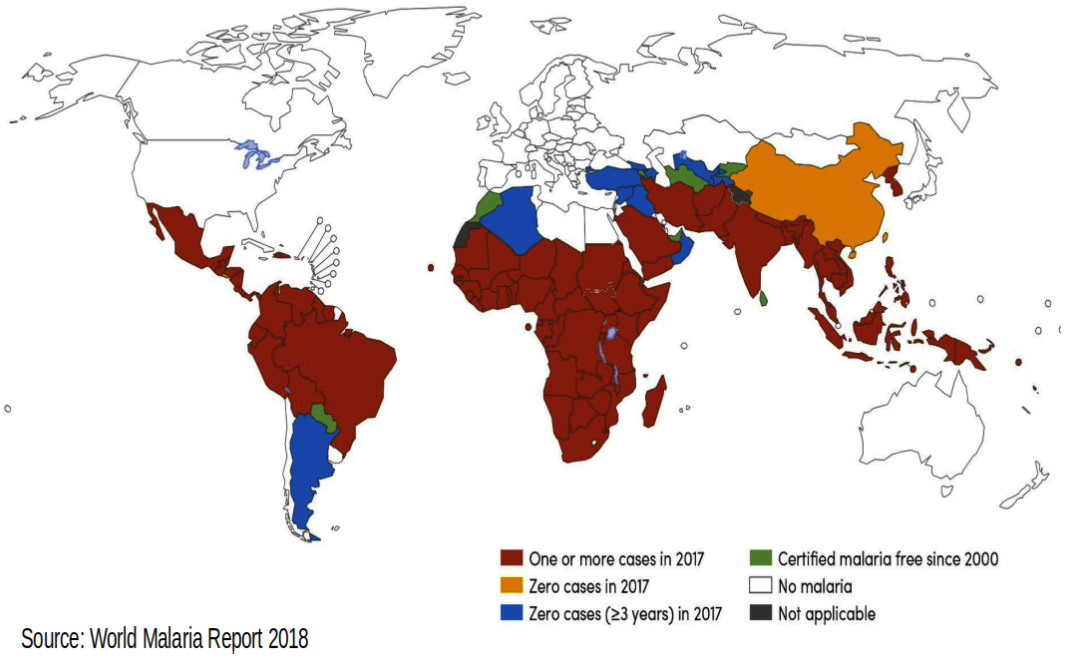

Figure 1. A diagram showing the global distribution of malaria.

Between the year 2000 and 2015, malaria occurrence rate reduced significantly by 37% globally and 42% in Africa with a corresponding decrease in mortality by 60% and 66% globally and in Africa respectively [4]. However, 2017 World Health Organization (WHO) malaria statistics showed that there were about 216 million global reported cases and 445,000 related deaths out of which 90% of both malaria occurrence and death occurred in sub-Saharan Africa (SSA) [5].
Mortality rate of the disease in areas with high transmission is mostly observed among children under five years and pregnant women [5]. Children are also at high risk of contracting the disease because they have not yet developed immunity. On the other hand, pregnant women are prone to malaria infection due to their modulated immune response and the accumulation of parasites in the placenta. In addition, the fetus is also at high risk because the parasites in the placenta could cause intrauterine growth restriction, resulting in cases such as low birth weight, congenital malaria and stillbirth [6].

There are five reported species of *Plasmodium* genus that cause human malaria. The species are *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale* and *Plasmodium knowlesi* [5, 7]. *Plasmodium falciparum* is ubiquitous mostly in the tropical Africa, Southeast Asia and the Amazon region [8]. It is the most virulent parasite associated with severe clinical symptoms and mortality, particularly in SSA, a region with about 3.2 billion people at risk [9]. Studies have shown that the genetic variation in *Plasmodium falciparum (P. falciparum)* populations is a function of geographic distance from SSA [10]. These variations significantly influences the pathogen's ability to evade human immune response.
P. falciparum malaria infection is the third most life-threatening infectious disease of humans and is associated with serious complications. These complications are as a result of parasite-infected red blood cells sequester in the microvasculature of organs [11]. *Plasmodium vivax* causes chronic disease and is the second most harmful human malaria causing species responsible for significant morbidity outside SSA, particularly in South America and the Asia-Pacific region, contributing about 16 million cases annually [5]. *Plasmodium vivax* presents specific biological challenges that make it difficult to detect its infections. This is because of its ability to remain hidden and dormant in an infected person's liver [3]. *Plasmodium ovale* like *Plasmodium vivax* have the added deadly complication of a dormant liver stage, which can be reactivated in the absence of a mosquito bite. It is reported that *Plasmodium ovale* and *Plasmodium malariae* represent only small percentage of malaria infections. *Plasmodium knowlesi*, the fifth human malaria parasite infection is prevalent in South-east Asian countries. *Plasmodium knowlesi* malaria infection could be transmitted from monkeys to humans as well as through *Anopheles balabacencis* [12]. Studies have shown it to be responsible for up to 80% of malaria infections in Malaysia [12].

Plasmodium falciparum has 23-megabase, (A+T) rich genomic content which encodes about 5,300 genes located on 14 chromosomes, along with 35-kb circular plastid genome and a 6kb mitochondrial genome [13]. The majority of its predicted proteome are cell adhesion and host immune system evasion proteins whiles enzymes and transporters constitutes the minority [14]. Unique survival mechanism of this parasite is its ability to evade host immune response by switching its variant surface antigens. These surface antigens mediate the binding of infected erythrocytes to the vascular endothelium (cytoadherence) and non-infected erythrocytes (rosetting). This results in the accumulation of infected cells in the vasculature of a variety of organs, blocking the blood flow and reducing the oxygen supply [15].

With no vaccine available yet in the market, malaria treatment and control is mainly by chemoprevention, vector control through the use of mosquito treated nets and insecticide sprays, together with diagnostic-led case management [16]. However, progress in malaria elimination is threatened by challenges including, but not limited to: rapid emergence of resistant strains towards chemoprevention and poor responses of vectors to insecticides and treated nets.
Resistance to antimalarial drugs by *P. falciparum* due to selective pressures has been a major public health burden specifically in SSA since the emergence of chloroquine (CQ) and sulfadoxine-pyrimethamine (SP) resistance. This challenge hinders the ability to effectively suppress the infection

due to the significant efficacy decline of antimalarial drugs.
The discovery of antimalarials especially chloroquine, helped to control and eradicate malaria infection globally through the National Malaria Control and Eradication Program in the 1950s until resistance emerged [17]. Drug resistance remains a major obstacle to malaria control and elimination globally.

Primarily, antimalarial drug resistance development emerges as a result of variations affecting the structure and molecular mechanisms of the drug target in the pathogen or affecting the ability of the drug to access its target to execute the expected biological activity. Substandard medications and high drug residual levels after treatment is a contributing factor that increases selective pressure on parasite populations thus enabling a conducive environment for spread of drug-resistant parasites. Drug efficacy and resistance emergence is determined by, but not limited to: clinical or *in vivo* efficacy trials, pharmacological studies, *in vitro* parasite susceptibility assays, *in vitro* drug susceptibility test, molecular studies, genetic cross-linkage and mapping studies of parasites [17]. Drug pressure can be used to identify loci responsible for drug resistance.

Although significant improvement over the past decade has helped to reduce malaria mortality and morbidity, resistance to antimalarial drugs have intensified the morbidity and mortality rate thereby hindering the progress of controlling, eradicating and eliminating the disease. This is because, resistance increases the risk of treatment failure thus resulting in severe malaria, anaemia and recurrent parasitaemia [18]. These resistances have been reported for *Plasmodium falciparum, Plasmodium vivax* and *Plasmodium malariae*. Among the three, *Plasmodium falciparum* have developed resistance to most antimalarial drugs and this poses a great threat to controlling and eliminating malaria globally. As such, identifying and understanding mechanisms of antimalarial drugs would contribute immensely to elucidate the patterns of emergence and the continuous spread of drug resistance within particular populations, thus, helping to put up strategic and effective policies to control the effect.

Multiple variations within a gene could result in differential contributions to drug resistance or susceptibility. Understanding molecular markers associated with malaria drug resistance is of noteworthy importance and would contribute immensely in monitoring the spread of resistant parasites by either measuring parasite responses to specific drugs or measuring the prevalence of specific variants associated with reduced drug efficacy [19]. Among the many molecular markers for drug resistance, *Plasmodium falciparum* ATP-binding cassette (ABC) transporters are known to be significantly involved in malaria drug resistance. Because of their transport mechanism across extra and intracellular membranes, variations and/or overexpression could result in drug resistance and treatment failure [20].

Extensive adherence and use of monotherapy in the past resulted in rapid resistance development by the parasite to most available antimalarials (**Table 1** shows available antimalarial drugs). After the development of resistance to quinine, chloroquine, sulfadoxine-pyrimethamine and mefloquine, artemisinin combination therapies (ACTs) have been adopted as first line treatment for uncomplicated malaria in most endemic countries especially SSA since early 2000s. This is because of its faster clearance rate and high efficacy [21]. Today, it is considered as the defense mechanism for malaria control. However, the advent of *Plasmodium falciparum* resistant strains are associated with reduced sensitivity to artemisinin and all the quinoline partner drugs in use including amodiaquine, lumefantrine and mefloquine [1]. In addition, evidence abounds concerning the spread of multi-drug resistance by the parasites from countries of the Greater Mekong Subregion, i.e., Cambodia, Thailand, Lao People's Democratic Republic, Myanmar, Viet Nam and Myanmar-China-India border [22]. Because of that, the drug is seen as a fading hope. Spread of resistant parasite strains is a major factor to the global resurgence of malaria research. Spread of resistance is linked to the fact that, the presence of drug within the parasite confers survival advantage which results in the transmission of resistant parasite [23]. These factors highly correlates with global malaria burden. Tolerance by resistant strains are characterized by a delayed parasitic clearance and a high rate of parasites recrudescence in populations under artemisinin selective pressure [9]. It has been reported that antimalarial drug resis-

tance is also compounded by cross resistance [23], a phenomenon where by one drug maybe selected for by another in situations where mechanism of resistance is similar.

Emergence of *Plasmodium falciparum* CQ and SP resistant strains in Africa is traced mainly to Cambodia where there are several reports of artemisinin resistance [24]. The current rapid spread of artemisinin resistance in SEA coupled with the gene flow of chloroquine and SP resistant parasites into SSA and its impact presents a serious concern of the likelihood of emergence and spread of artemisinin resistant strains in Africa [22].

To date, there is no confirmed report of ACTs resistance in Africa, South America and Oceania [9]. Researchers have proposed that with the rising trend in malaria cases and death in Africa, should resistance to ACTs (particularly artemisinin) emerge, there would be about 78 million additional malaria cases which would result in an increase in the morbidity and mortality rate between the year 2016 and 2020 [9].

With the aforementioned challenges faced by malaria treatment and control, the need to understand antimalarial drug resistance in Africa and identify novel, safe and effective antimalarial chemotypes specific to the African populations will persist until the human pathogenic *Plasmodium* species are eventually eradicated. This research proposes to understand the mechanisms and patterns underlying antimalarial drug failure in SSA by investigating functional interactions between malaria selective variants using bioinformatics pipelines and techniques. The study proposes also to identify potential protein targets that can play their role in predicting potential drug candidates specific to the African populations through a comparative approach of both host and *Plasmodium falciparum* genomics analysis.

1.2 Life cycle of *Plasmodium falciparum*

The life cycle of *Plasmodium falciparum* involves different hosts and tissues as shown in **Figure 2**. The cycle comprises of three different developmental stages: the mosquito stage, the human liver and blood stages.

Malaria parasites alternate between asexual replication in the bloodstream of human host and reproduction in the vector. Asexual replication causes structural and functional changes which results in the disease, while reproduction in the vector facilitates transmission to new hosts [25]. Clinical symptoms of malaria are observed during the parasite's intra-erythrocytic developmental stage inside the host's red blood cell. The life cycle of the parasite is quite complex, and as such the pathogen establishes molecular mechanisms to ensure its growth and survival within its hosts.

The ability of the parasite to survive and develop within the environments of its hosts is facilitated by genes and their specialized proteins that interact with the host cell to modulate the host's response to it [26]. These interactions help the parasite to invade and grow within multiple cell types as well as evade the human immune system.

Human infection, termed as blood meal, begins when malaria infected female *Anopheles* mosquitoes inoculates sporozoites from their salivary glands into the host. These sporozoites migrate into the bloodstream and into the liver, where they begin the liver stage [27]. The sporozoites invade hepatocytes and mature into schizonts. The schizonts undergo dormancy then after replicating, they rupture to release merozoites [28].

The merozoites undergo asexual multiplication and then invade human mature red blood cells (RBCs) to begin the blood stage. They develop inside a parasitophorous vacuole and initiate almost a 48 hour cycle of asexual blood stage (ABS) parasite growth, egress and re-invasion [29]. The ring stage trophozoites mature into schizonts, which rupture to release merozoites. Some of the parasites differentiate into sexual gametocytes (erythrocytic stage). Blood stage parasites are responsible for the clinical manifestations of the disease. The gametocytes, male (microgametocytes) and female (macrogametocytes), are ingested by an *Anopheles* mosquito during a blood meal. It is estimated that about 10000–100000 mature gametocytes are taken up during a blood meal. These gametocytes then

form male and female gametes that undergo sexual recombination to form ookinetes and then oocysts before completing their life cycle by forming sporozoites that migrate to salivary glands of the vector ready for further human infection. Inoculation of the sporozoites into a new human host perpetuates the malaria life cycle.

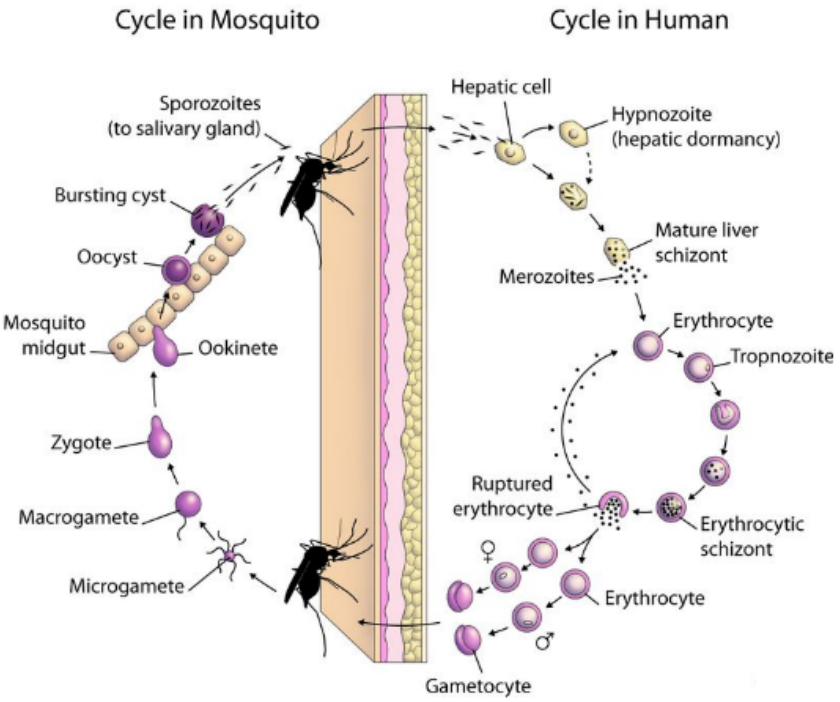

Figure 2. Schematic diagram of the life cycle of *Plasmodium falciparum* in both human and mosquito. Image retrieved from Malaria Site (https://www.malariasite.com/life-cycle/).

1.3 Antimalarial drug candidates

Table 1. Description of various antimalarial drug candidates.

Antimalarial derivatives	Chemical family	Drug name	Use	Antimalarial activity	Reference
Quinoline derivatives	4-aminoquinolines	Chloroquine	Treatment of both *falciparum* and non-falciparum malaria infection	Inhibits heam polymerization	[30]
		Amodiaquine	Treatment of non-severe *falciparum* infection	inhibits heam polymerization	[17]
	Bisquinoline	Piperaquine	Treatment of both *falciparum* and *vivax* malaria infection	inhibiting detoxification of haem	[31]
	Arylamine alcohols	Quinine	Treatment of severe *falciparum* malaria infection	Accumulates in food vacuoles and forms toxic haem complexes	[17]
	Arylamine alcohols	Mefloquine	Treatment of non severe *falciparum* malaria infection. Antimalarial prophylaxis	inhibits digestion of haemoglobin	[17]
		Pyronaridine	Treatment of non severe *falciparum* malaria infection		[32, 33]
		Naphthoquine	Used in combination with artemisinin to treat both *P. falciparum* and *P. vivax* malaria infection in children		[34]
Phenanthrenes and derivatives	Arylamine alcohols	Halofantrine	Treatment of non severe *falciparum* malaria infection	Causes parasite membrane damage by forming cytotoxic complexes	[17]

Continued on next page

Table 1 – continued from previous page

Antimalarial derivatives	Chemical family	Drug name	Use	Anti-malarial activity	Reference
		Lumefantrine	Treatment of non severe *falciparum* malaria infection		[17]
	8-aminoquinolines	Primaquine	Treatment of *P. vivax* and *P. ovale* malaria infection	Blocks oxidative metabolism in the parasite	[35, 36]
	8-aminoquinolines	Tafenoquine	Treatment of *P. vivax* and *P. ovale* malaria infection	inhibits haem polymerization	[37]
	8-aminoquinolines	Quinidine	Treatment of *Plasmodium falciparum* malaria infection	Accumulates in the food vacoule and forms toxic haem complexes	[38]
	Benzene and substituted derivatives	Sulfamethoxypyridazine		Inhibit synthesis of folates	[39]
Benzene and substituted derivatives	Antifolate	Sulfadoxine		Inhibits folates synthesis	[39]
Diazines	Antifolate	Pyrimethamine	Treatment of *Plasmodium falciparum* malaria infection	Inhibits enzymes within the folate pathway	[16]
	Antibiotic	Doxycycline		Inhibits protein synthesis	[40]
Carboxylic acids and derivatives	Antibiotic	Clindamycin	Complicated *falciparum* malaria	Inhibits protein synthesis	[41, 42]
hydroxy-naphthoquinone		Atovaquone	Treatment of non severe *falciparum* malaria infection	Affects mitochondrial electron transport chain	[43]

Continued on next page

Table 1 – continued from previous page

Antimalarial derivatives	Chemical family	Drug name	Use	Anti-malarial activity	Reference
Biguanide	Benzene and substituted derivatives	Proguanil		Inhibit synthesis of folates	[43]
Artemisinin derivatives		Artesunate	Treatment of non severe *falciparum* malaria infection	Affects mitochondrial electron transport chain Disrupts redox cycling Inhibits haem metabolism	[17]
		Artemether	Treatment of non severe *falciparum* malaria infection	Affects mitochondrial electron transport chain Disrupts redox cycling Inhibits haem metabolism	[17]
		Dihydro-artesunate	Treatment of non severe *falciparum* malaria infection	Affects mitochondrial electron transport chain Disrupts redox cycling Inhibits haem metabolism	[17]
		Arteether	Treatment of non severe *falciparum* malaria infection	Affects mitochondrial electron transport chain Disrupts redox cycling Inhibits haem metabolism	[31]

1.4 Antimalarial drug Response/Resistance and *Plasmodium falciparum*

The genetics of the parasite constitute a major building block that determines the variations in the levels of drug susceptibility, particularly having elucidated that antimalarial drug resistance of *P. falciparum* involves a single major gene effect [17]. These variants could be either single or multiple point which arises from a series of linked or unlinked additive variations. This resistance is effectuated by resistance-associated variants under selection and the geographical spread of resistance alleles from their origin [17]. **Table 2** describes *Plasmodium falciparum* selective genes and mutations associated with drug resistance.

Table 2. Description of genes/selective variants associated to various antimalarial drug resistance.

Drug	Gene	Variation	Reference
Chloroquine	*pfcrt*	*K76T, K76N, K76I, C72R, S163R*	[44]
	pfmdr1	*N86Y, D1246Y, S1034C, N1042D*	[45]
	pfmrp1	*F1390I, Y191H, A437S*	[17]
Pyrimethamine	*pfdhfr*	*S108N/T, N51I, I164L, A16V, C59R*	[46, 47]
Sulfadoxine	*dhps*	*S436A/F, A437G, K540E, A581G, A613S/T*	[47]
Piperaquine	*plasmepsin 2–3*		[48]
	pfmrp1		[49]
	pfcrt	*F145I*	[50, 51]
	pfk13		[50]
	pfmdr1	*N86Y*	[52]
Mefloquine	*pfmdr1*		[53]
Artemisinin	*pfatp6*	*L263D,L263E, L263K*	[54, 55]
	pfk13	*C580Y, Y493H, R539T, I543T*	[17]
	pfcrt		[1, 17]

1.4.1 Resistance/response to chloroquine

The discovery of chloroquine, a rapidly-acting schizonticidal drug, in the 1940s served as the primary chemotherapeutic means of malaria treatment because of its efficacy, safety and low cost. This tremendous breakthrough helped reduce reported malaria cases and mortality rate.
It's usage contributed immensely to controlling malaria until the emergence and spread of drug-resistant *Plasmodium falciparum* strains two decades after its introduction, with the first reports from

Thai-Cambodia border and Columbia followed by South America [56, 57]. Resistance to chloroquine is characterized by diminished accumulation, decreased sensitivity and reduced concentration within the parasite's food vacuole. The lack of new antimalarial drugs led to the spread of chloroquine resistant parasites to Africa around 1970s which resulted into about three fold increase in the mortality rate [17].

Free haem produced from heamoglobin digestion is suggested as target for chloroquine [58]. The mode of action of this 4-aminoquinoline compound involves accumulation inside the digestive vacuole of the intraerythrocytic trophozoite (infected red blood cell) where its concentration increases [8]. It then inhibits haemoglobin degradation and binds to haem moiety to form lethal haem-chloroquine complexes. The complexes formed inhibits haem polymerization, a process that detoxifies haem, thus, inhibiting the production of haemazoin pigment catalyzed by *Plasmodium falciparum* histidine-rich protein-2 (*Pfhrp-2*) [58]. This inhibition allows the accumulation of toxic hemoglobin metabolite in the cell thereby disrupting the biochemical processes of the parasite and leading to cell death [59]. A study conducted by Tewari and colleagues to investigate the effect of chloroquine on *P. falciparum* DNA replication identified significant changes in DNA synthesis related genes during chloroquine treatment. The study reported continuous accumulation of haem within the parasite's food vacuole to be associated with inhibition of redox metabolism, carbon fixation and pyrimidine metabolism which are contributing factors of DNA synthesis inhibition [60].

Molecular marker analysis, genetic mapping and allelic association studies have shown that chloroquine resistant *P. falciparum* strains emanated from Asia through South America and to Africa due to variations in *P. falciparum* chloroquine resistance transporter (*pfcrt*) gene on chromosome 7 [30, 44]. It is well established that the spread of chloroquine resistant parasite strains through multiple evolutionary pathways is associated with *pfcrt K76T* variant [56]. However, other variants in the transmembrane domains of *pfcrt* encoded protein, specifically, *K76N, K76I* and *C72R* in transmembrane 1, *S163R* in transmembrane 4, and *Q352K* and *Q352R* in transmembrane 9 have reportedly been linked to chloroquine resistance in clinical studies and field isolates [19]. The parasite's ability to remove chloroquine from its food vacuole is suggested to depend on the *K76T* variant because it is consistent in chloroquine resistant isolates studied.
N86Y, D1246Y, S1034C and *N1042D* variants and copy number variants in the *Pfmdr1* gene on chromosome 5, which codes for a homologue of human multi-drug resistance p-glycoprotein (Pgh1) protein in the digestive vacuole of *Plasmodium falciparum*, is associated to chloroquine reduced susceptibility and resistance [45]. Also, *Pfmrp1* gene variants including *F1390I, Y191H* and *A437S* have been reported to be linked with in vitro susceptibility to chloroquine [17].

Several studies have reported variations in the *pfcrt* gene in individuals of Asian, African or South American origin that confer verapamil-reversible chloroquine resistance, cross-resistance, as well as alter susceptibility to other anti-malarial drugs such as quinine and quinidine which targets the parasite's food vacuole [61]. Also, different phenotypic expressions by resistant isolates that carry the *pfcrt* and *pfmdr*1 alleles give an indication that there are other genes responsible for modulating the pathogen's response to drugs [13]. These modulators could be molecules involved in drug transport or transporter genes encoding products like ABC transporters.
The advent of chloroquine resistant parasites with adequate fitness to survive and spread over time resulted in the decline of its efficacy. This resulted to the abandonment of the drug in early 2000s [17]. It was used for about 50 years in Africa before it was withdrawn from the market.

1.4.2 Resistance/response to sulfadoxine – pyrimethamine (SP)

The widespread of CQ resistance led to the introduction of SP as an alternative antimalarial drug to treat CQ resistant *falciparum* malaria. Sulfadoxine and pyrimethamine act with high synergistic effect when administered together to inhibit dihydrofolate reductase and dihydropteroate synthetase which are two enzymes important in the parasite's folate biosynthesis pathway [62, 63]. SP binds to these

enzymes to inhibit their activity. This inhibition results in significant decrease of tetrahydrofolate produced which sequentially results in a reduced production of folate precursors such as methionine and deoxythymidine monophosphate (dTMP). Low production of folate precursors inhibits the parasite's life cycle [47]. In the early 1980s, SP was adopted in Africa for treating non-severe malaria. SP is used either as monotherapy or in combination with other antimalarial agents such as artemisinin-based derivatives to treat uncomplicated *Plasmodium falciparum* malaria as well as for intermittent preventive therapy in infants and pregnant women in Africa [17]. Clinical efficacy trials have shown that SP intermittent preventive therapy confer protection against malaria anemia in children and improve foetal outcomes from malaria [47].

The extensive use of SP led to a rapid resistance and decrease in efficacy during its year of introduction [17]. The resistance to SP is influenced by the influx of folates and the efflux of antifolates during the biosynthesis. The prevalence of SP resistant parasite genotypes was first reported in Thailand before spreading to other malaria endemic regions. Genetic crosslinking and mapping studies helped to identify variations involved in SP resistance. Evidence abounds for selective sweep of these genetic variations into Africa [64].

The resistance to pyrimethamine is linked to variations in the parasite's dihydrofolate reductase gene (*pfdhfr*) on chromosome 4 [65]. *Pfdhfr* variants *S108N/T, N51I, I164L, A16V* and *C59R* results in an increased parasite clearance [46, 47]. Single nucleotide variants such as *S108N* expresses low resistance whiles *N51I/S108N* and *C59R/S108N* double-variant expresses relatively higher resistances. *S108N, N51I* and *C59R* is the most common triple-variant mutant characterized by increased rates of treatment failure in high pyrimethamine resistance areas in Africa and Southeast Asia [46, 64]. These triple-variants increase and diverge rapidly in some parts of Africa as compared to other variants. Among the variations, *N51I /C59R/S108N/I164L*, a quadruple-variant, highly prevalent in Southeast Asia but with low frequency in Africa, is considered to confer high resistance to pyrimethamine such that it renders the parasite untreatable [46, 66]. Microsatellite analysis have shown that these triple-variants have shared ancestry with *pfdhfr* variants in Southeast Asia thus, suggesting the spread of pyrimethamine resistance from Southeast Asia to Africa just as the emergence of chloroquine resistance [66]. Similarly, the *pfdhfr* double-variant in Africa emerged from three independent origins whereas the triple-variant occurred from a single origin thus confirming the spread of single dominant resistant parasite lineage into Africa [46]. However, there are additional *pfdhfr* haplotypes present in Africa at low frequencies particularly in Kenya and Cameroon [67]. The *C50R* variant of *pfdhfr* in South America has been detected in Africa and thought to increase the level of resistance [68].

On the other hand, selection of parasite with variations in dihydropteroate synthetase gene (*pfdhps*) on chromosome 8 results in resistance to sulfadoxine [65]. *S436A/F, A437G, K540E, A581G and A613S/T* variants in *pfdhps* confer reduced susceptibility to sulfadoxine and daspone, another antifolate drug [47]. Studies have shown that these variants express reduced affinity for sulfadoxine. Double-variants *A437G* and *K540E* have been shown to be very common in Africa. Studies have shown that *S436A* and *A437G* alone express low resistance level but they act synergistically with *K540E*, *A581G*, and/or *A613S* to confer an increased level of resistance to sulfadoxine [64]. These mutations results in a long half-life of the drug. Other studies have shown that selection of parasites with highly resistant allelic types also confers resistance [63].

The level of SP resistance is determined by the type of variant. Single nucleotide variants result in about 100-fold lower levels of resistance compared to wild-types, while multiple variations results in about 225-fold higher level of resistance compared to wild-types.
A study conducted in 31 African countries revealed consistency in the reduced efficacy of SP due to significant increase in *pfdhfr* triple-variant and *pfdhps* double-variant genotypes, with Kenya-Tanzania border and Malawi recording high prevalence rate [47]. Natural selection and hard selective sweeps contributes to the gene flow of resistant alleles and genetic hitchhiking across the parasites population in Africa [69]. Although there is independent segregation of *pfdhfr* and *pfdhps* alleles in Africa, there is lack of variation and significant linkage disequilibrium between codons 51, 59, and 108 in *pfdhfr*

and codon 437 of *pfdhps* due to drug selection [64].

1.4.3 Resistance/response to piperaquine

Piperaquine is characterized mainly by its large distribution volume, ability to bind high number of plasma proteins and long half-life resulting in low hepatic elimination clearance [70]. The uncontrolled use of piperaquine as monotherapy in China in the 1970s and 1980s contributed significantly to the development of parasite resistance, which were validated through several clinical drug resistance and *in vitro* reports [31]. Aside from that, cross-resistance between piperaquine and other antimalarial agents including artemisinin derivates and chloroquine could be a factor. The long half-life of piperaquine causes it to circulate within the body for long, and puts other partner drugs (in combinatorial therapy) at high risk of selection.
The degree of cross-resistance varies among other antimalarial agents. For instance, an *in vitro* test on two piperaquine resistant strains revealed a cross-resistance between piperaquine and hydroxypiperaquine which is formed from oxidation of piperaquine. Also, a cross-resistance has been reported between piperaquine, artesunate and mefloquine [31]. Among other strains, moderate cross-resistance was observed between piperaquine and pyronaridine. A recent study conducted by Witkowski et al. [48] in western Cambodia showed that piperaquine resistance in Cambodia is strongly associated with amplified copy number in the *plasmepsin 2–3* gene on chromosome 14, which encodes haemoglobin-digesting proteases. These enzymes have been shown to modulate parasite susceptibility to piperaquine [71]. Combined analysis of K13 polymorphisms and plasmepsin 2 copy number have been identified as signature of selection for dihydroartemisinin–piperaquine failures. A variant in the *pfmrp1* gene is reported to be associated with piperaquine resistance [49]. In a recent study to determine piperaquine sensitivity and the role played by *pfmdr1*, it was shown that selection on *pfmdr1* 86Y allele was associated with reduced piperaquine sensitivity in an *in vitro* test using Thai *Plasmodium falciparum* isolates [52]. Agrawal and colleagues in a genome-wide association study identified *F145I* as a novel variant of the *pfcrt* gene linked with reduced sensitivity thus resulting in dihydroartemisinin-piperaquine treatment failure [51]. Duru and colleagues, in a study in Cambodia confirmed that *Pfmdr1, Pfcrt* and *Pfk13* variants are associated with piperaquine resistance [50].

1.4.4 Resistance/response to mefloquine

Mefloquine monotherapy was mostly used in South-east Asia until the emergence of parasite resistance rendering it ineffective within six years after its introduction as an antimalarial drug in 1984. However, it is now used in combination with artemisinin and has been shown to be effective. The exact mechanism of resistance is unknown, however, there are several lines of *in vitro* and *in vivo* experimental evidences that explicitly show that an increase in copy number of multidrug-resistant gene 1 (*pfmdr1*) is associated with mefloquine resistance [53]. Studies conducted by Price et al. showed that increased copy number in *pfmdr1* is associated with attributable hazard ratio for treatment failure of 6.3 and 5.4 for monotherapy and combination therapy respectively [72]. *In vitro* studies showed an increase in copy number of *pfmrd1* was associated with a significant decrease in susceptibility to mefloquine. In their study on *Plasmodium falciparum* isolates from Suriname, Labadie-Bracho et al. showed that an increase in copy number of *pfmdr1* is not only associated with mefloquine resistance but also with resistance artemether-lumefantrine *in vitro* [53]. Amplification of *pfmdr1* was reported to be associated with an increase in mefloquine IC50 [13]. The gene *pfmdr1* has therefore emerged as a critical determinant associated with many malaria drugs resistance by modulating their sensitivity. This phenomenon could result from an extensive selection of multiple drugs on the parasite. Also, Preechapornkul et al. studied the ex-vivo dynamics of *pfmdr1* from a Thai *Plasmodium falciparum* isolate using multiple genetic markers and concluded that *pfmdr1* amplification is obtained as a result of mefloquine resistance in addition to the decline in the parasite's survival fitness with no drug pressure [73].

1.4.5 Resistance/response to ACTs

Further research helped to implement the policy of using antimalarial drugs as combination therapies instead of the usual monotherapy. This impelled the introduction of Artemisinin Combination Therapies (ACTs), which has demonstrated tremendous positive therapeutic response. This is because, the components of the drug have different modes of action that enhance pharmacodynamic synergistic effect whiles significantly minimizing resistance development of the parasite by decreasing the risk of selection by resistant parasites compared to a monotherapy setting [31].
In these drugs, artemisinin derivates sesquiterpenes found mainly in plants of the genus Artemesia, are used as the core component in combination with other antimalarial drug agents thereby preventing recrudescent malaria [31]. Due to the short half-life of artemisinin, it permits a very sharp reduction of parasite load in an infected individual. The efficacy of artemisinin is linked to the fact that they target both the early and late erythrocytic stages of the parasite where it kills the parasite through an induced proteopathy mechanism or degeneration of parasite's cytoplasm [74]. Only residual parasites are then eradicated by the partner drugs which are characterized by relatively increased half-life as compared to artemisinin [31].
ACTs recommended by WHO are artesunate and mefloquine; artesunate and amodiaquine; artesunate, sulfadoxine, and pyrimethamine; artemether and lumefantrine; and dihydroartemisinin and piperaquine [55]. In Africa, artemether-lumefantrine (AL) and artesunate-amodiaquine are mostly used in endemic areas because they show the highest efficacy whereas dihydroartemisinin-piperaquine (DP) is used mostly in Asia. Artesunate-mefloquine is shown to be effective and safe for treating malaria among children below 5 years of age in Africa [75].
Artemisinin reacts with heam to form heam-artemisinin adducts in the parasite's food vacuole. These adducts interact with proteins in the parasite responsible for heam detoxification and inhibit haemozoin polymerization. This interaction leads to accumulation of toxic heam which results in parasite's damage [9]. Artemisinin kills the young intraerythrocytic malaria parasites by inhibiting the parasite's enzyme (PfATP6) encoded by the *pfatp6*, a calcium transporting ATPase gene [55].

The use of combination therapies is to increase drug efficacy and to delay parasite resistance [17]. However, there is a high risk of tolerance or resistance to ACT by *Plasmodium falciparum* especially in endemic areas. This phenomenon is characterized by reduced ring stage susceptibility, prolonged parasite clearance and parasite recrudescence as a result of selection due to artemisinin partner drugs which circulate longer at decreased levels [76]. Artemisinin resistance therefore contributes significantly to parasite selection to partner drugs [22]. However, in Africa, amodiaquine and lumefantrine are of much a concern since resistance to them is prevalent. This phenomenon, therefore increases the chance for resistance emergence to artemisinin derivatives. Nevertheless, there are reports of resistance to artemisinin, an integral component of ACTs, apart from its first report in Cambodia [24]. *P. falciparum* resistance to artemisinin is rapidly evolving in the Greater Mekong Subregion through selective sweeps [22]. This resistance leads to a significant decrease in the parasite biomass clearance rate, thus, resulting in increased load of residual parasites unable to be cleared by partner drugs. Subsequently, this phenomenon brings about increase in recrudescent and spread of resistance.

Genome analysis has revealed that variation in the propeller domain of the gene encoding Kelch protein 13, such as *C580Y, Y493H, R539T* and *I543T* are associated with higher artemisinin resistance [17]. These variations have been studied to have emerged independently from multiple occasions [77] and confer resistance. Recent studies have shown that low levels of immunity to *Plasmodium falciparum* in the Mekong population in South-east Asia is associated with high prevalence of variations in the propeller domain of Kelch protein 13 encoding gene (*kelch13*) [9]. The variant *PfKelch13 C580Y* in parasite lineage observed in Cambodia, northeastern Thailand and southern Laos is reported to be the most dominant and to be more transmissible such that it has now spread to 3 countries in the Mekong region thus causing high ACT failure rate [22]. *Pfplasmepsin2* gene amplification has emerged on the back of *PfKelch13 C580Y* variant and contribute to the widespread of multidrug-resistant parasite lineage [22]. Studies have shown that the *Pfplasmesin2* gene, involved

in hemoglobin digestion pathway, is targeted by artemisinin. It is suggested that *Pfplasmesin2* amplified parasites evolved after selection of the *pfkelch13 C580Y* lineage, which has intensified ACT treatment failure. In a study to assess prevalence of kelch13 polymorphisms within Sub-Saharan Africa, 22 unique variants comprising of 7 nonsynonymous SNPs were identified. *A578S* and *V566I* variants with frequency > 1% were found to be present in 5 African countries. They are closer to the *C580Y* variant but confer no resistance to artemisinin [78]. There is, therefore, a higher possibility that the K13 variant that confer resistance work together with other factors depending on the parasite population. Interestingly, the K13-propeller *M476I* together with the $PF3D7\_0110400D56V$ variant have been reported to confer resistance to artemisinin in the artemisinin-resistant parasite line from Africa [79].

Alterations in the parasitic membrane proteins Pgh-1, PfCRT and PfMRP1 are major contributors to artemisinin resistance through decreasing intracellular drug accumulation by their active drug transport mechanism between the food vacuole lumen and the cytoplasm [1, 17]. A disruption of the *pfmrp1* gene within a parasite provides a higher sensitivity to artemisinin, chloroquine and lumefantrine, thus, confirming the gene's ability to influence the parasite's drug response [20]. A genome-wide association study using field isolates from China-Myanmar boarder, further implicated the autophagy-related protein 18 (PfATG18) to be associated with decreased sensitivity to artemether, dihydroartemisinin and piperaquine [16].

Research findings by Veiga et al. [1] on association between copy number variation of *pfmdr1* and drug sensitivity together with other studies highlighted observations that increase in copy number decreased sensitivity of *Plasmodium falciparum* to artemisinin, lumefantrine and mefloquine [55]. The *Pfmdr1* variants *N86Y, Y184F, D1246Y* and *N1042D* and the *pfcrt* 76T variant are shown to confer resistance to amodiaquine and lumefantrine [17, 80]. A recent study in eastern Africa reported no significant selection for *pfmdr1* polymorphisms upon treatment and re-treatment of children between 12 and 59 months with artemether-lumefantrine and artesunate-amodiaquine thus supporting the use of same or alternative ACTs for malaria treatment [80]. *In vitro* analysis revealed diminished response to artemisinin and lumefantrine with the *pfmdr1* wild-type alleles N86 and D1246 [81]. However, a study conducted in Uganda revealed that the *pfmdr1 86Y* and *1246Y* variants mediated diminished response to amodiaquine [76].

A recent study conducted in Senegal, Mali and The Gambia to determine the frequency of *pfcrt* and *pfmdr1* variant associated with artemether lumefantrine resistance reported a decrease in the frequency of the *pfcrt K746T* and the *pfmdr1* variants at codon 86 during the study period [19]. The study reported a decrease in parasite sensitivity to artesunate and lumefantrine in Senegal.

With the spread of multidrug-resistant parasites from Cambodia, southern Laos and northeastern Thailand, its emergence in Africa would be a major public health concern.

1.5 Influence of human genetic variations on antimalarial drug resistance

Resistance to antimalarials is complex and involves host genetics, environmental factors and parasite compensatory variants. Having established that malaria has the strongest evolutionary selection on the human genome, there is a higher probability of parasite genetic diversity influencing the host genome leading to significant changes in immunity, parasitemia and malaria risk [82]. There is no direct relationship between evolution of the human genome towards drug resistance as compared to the parasite's genome, however, the host genetic factors contribute significantly to drug metabolism and effectiveness. Variations in host genes involved in drug metabolism such as the cytochrome P450 genes may down-regulate processes including but not limited to oxidative processes, hydrolysis and hydrophilic functionalities leading to poor catalytic activities and renal excretion.

It is important to note that, acquired host immunity a dependent variable of population endemicity,

drug exposure and transmission intensity influences drug clearance rate [82]. High immunity would mean low disease infection rate, low drug pressure resulting in delayed drug resistance development and vice-versa. Recent genome-wide studies have reported significantly diminished protective and non-replication of some host protective variants in some population although the general frequency of resistant alleles is high [83]. For instance traditional HLA alleles such as HLA-B*53 associated with malaria resistance showed no evidence of association in Gambia population [83]. Also, there are reported complication of the nature of protection of malaria protective variants such as G6PD predominant in Africa. Moreover, new protective variants have been reported [83]. Because Africa is characterized with high levels of immunity and faster parasite clearance, early signs of low-grade drug resistance could go undetected and this presents a challenge in investigating early signs of resistance [82].

1.6 Problem statement

Genetic variations are believed to be the cause of *Plasmodium falciparum* drug resistance [2]. Genetic polymorphisms of candidate genes from the disease causing pathogen generally provide effects that counteract the drugs controlling the disease. In view of that observation, spontaneous alterations in the form of single nucleotide variation in different genes of the *Plasmodium falciparum* genome, enhances the pathogen's ability to develop strategic mechanisms to tolerate or resist the drug action over time, thus, yielding the unexpected result [8]. Drug resistance poses a major challenge to the quest to controlling, eliminating and eradicating malaria. It is a principal reason for the expansion of this life-threatening disease.

The use of antimalarial drugs has been the optimal avenue for malaria control and artemisinin-based combination therapies (ACTs), which are presently the first line of treatment are used globally [1]. ACTs were adopted in Africa after the decline in efficacy of previous widely used antimalarial drugs, including chloroquine and sulfadoxine–pyrimethamine. This was to ensure that, each component of the combinatorial drug acts through different mechanism within the parasite, with the aim to reduce the likelihood of emergence of multi-drug resistant parasites significantly.

Unfortunately, the *Plasmodium falciparum* parasite has shown a tremendous ability to develop resistance and tolerance to these artemisinin derivatives and to the long half-life ACT partner drugs in some countries of the Greater Mekong Sub-region [84]. Accordingly, there is a higher likelihood of the emergence and spread of artemisinin–resistant parasites in Africa just as it was the case for chloroquine and sulfadoxine–pyrimethamine resistant parasites. Several reports support the significant decrease in the therapeutic response of artemisinin derivatives against *P. falciparum* [9, 85]. There is evidence of the appearance of ACT resistant strains of *P. falciparum* in Africa where the disease–associated morbidity and mortality rate are highly significant [9, 57]. This raises questions about the mechanism of resistance and the efficacy of the drug in Africa and the hypothesis that the resistance and susceptibility dilemma could result from a selection by the parasite.

1.7 Need statement

The emergence of artemisinin resistance has led to the identification of novel therapeutic targets, candidates and small-molecule inhibitors. Due to the selection resulting in resistance development by *Plasmodium falciparum*, this research study proposes the need to clearly understand the mechanism of action of Artemisinin Combination Therapies (ACTs) antimalarial drug and assess the prevalence of drug resistance. Using bioinformatics analysis, this study is proposing to identify specific molecular markers under selection that are involved with drug resistance. Through gene mapping and protein-protein interaction networks between the parasite and human, this study will help elucidate the target mechanisms. This is subjected to the use of systematic approach through a large-scale data-driven integrative framework to identify targets and more effective drugs.

1.8 Aims and objectives

The overall aim of this study is to identify potential protein targets to understand the mechanism of drug resistance and identify those molecular targets that can help design drugs that are more effective and more specific to the African populations through a comparative analysis of both host and *Plasmodium falciparum* genome.

Specific Aim 1:

Leveraging malaria-specific genome-wide association studies(GWAS) summary statistics from African populations, *Plasmodium falciparum* selective pressure variants and functional dataset (host-pathogen intra-organism and host-pathogen inter-organism PPIs) will be used to construct overlapping networks for both host and pathogen.

a Identifying selective pressure variants from literature and databases to construct malaria specific subnetwork of the pathogen.

b Retrieving and integrating inter and intra functional protein-protein interaction data between host and pathogen into a unified framework.

c Constructing overlapping network between the host and the pathogen.

Specific Aim 2: Leveraging disease–target–drug relationships to identify protein targets from the overlapping functional network of host and pathogen.

a Integrating all generated subnetwork to identify densely connected subnetworks and hubs.

b Elucidating clusters or densely connected subnetworks containing important functional proteins by computing functional similarity score.

c Performing disease-associated process and pathway enrichment analysis to identify the involvement of proteins in essential processes.

d Drawing conclusion on the drug failure mechanisms and proposing drug target variants.

Specific Aim 3: Predict potential drug candidates for identified targets using semantic similarity approach.

a Investigating disease similarity by exploiting relationships between gene ontological terms and disease pathology.

b Identifying repurposable approved drug candidates by leveraging drug-target-disease associations and pharmaceutical datasets.

1.9 Methodology

The general schematic representation of the methodology for this study is shown in **Figure 3**.

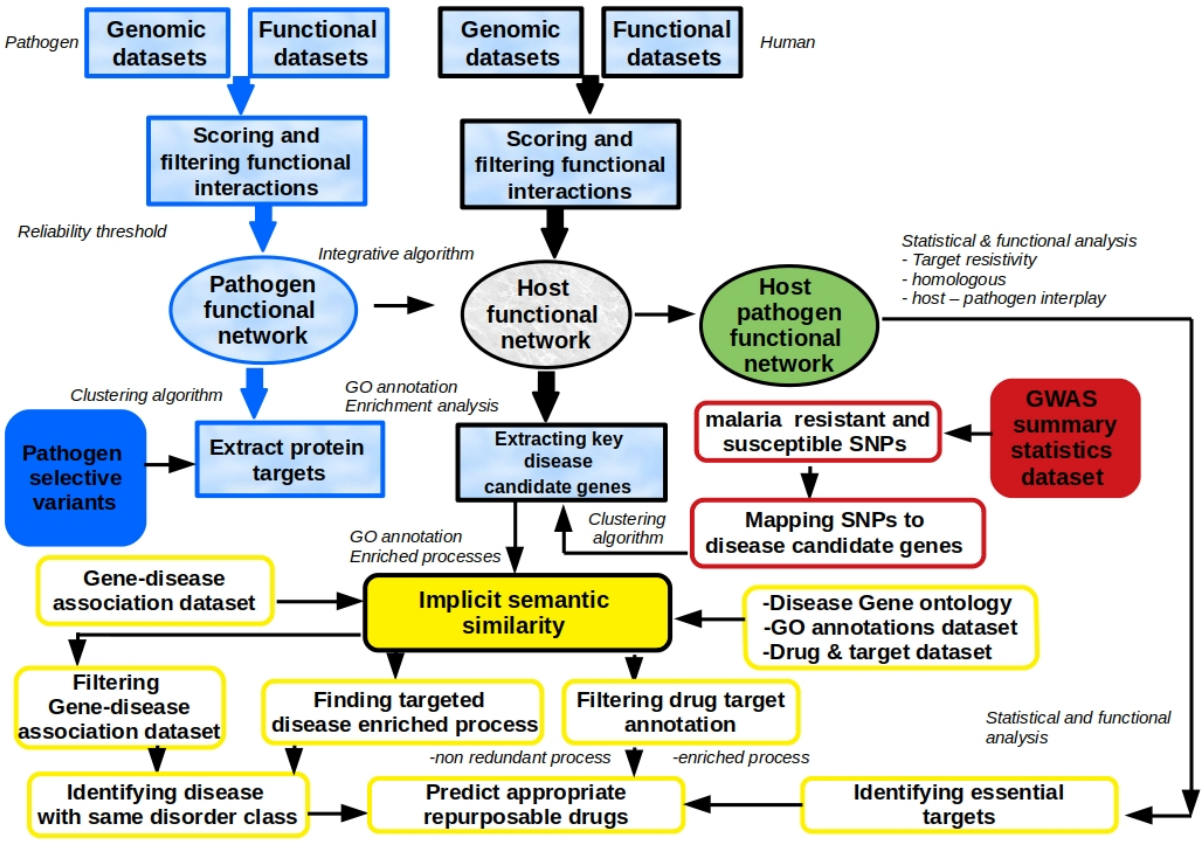

Figure 3. Diagrammatic representation of general methodology implemented in this study.

Specific Aim 1: Identifying susceptible and resistant human genes associated with malaria from previous Genome-wide association studies in African populations.

a. Host-malaria associated genes:

   This study proposes to use genome-wide association studies (GWAS) summary statistics shown in **Table 3** and bioinformatics approaches to identify and investigate genetic polymorphisms in selected African populations that are subjected to a high burden of malaria susceptibility and resistance. Disease-associated genes identified from analysis of GWAS summary statistics data will be used to construct malaria–specific subnetwork using bioinformatics tools.

   This approach will facilitate assessing malaria-associated genes in specific African population thus enhancing novel target identification for drug discovery.

Table 3. MalariaGen Human GWAS Data set on which the analysis will be conducted retrieved from www.malariagen.net/data/human-gwas/access-apply.

Datasets	Sample size	EGA Dataset ID	Ref.
Genome-wide study of resistance to severe Malaria in eleven populations (Kenya)	1944 cases, 1708 controls, 180 parents & 33 other	EGAD00010000904	[86]
Genome-wide study of resistance to severe Malaria in eleven populations (Gambia)	2807 cases, 2786 controls & 1 parent	EGAD00010000902	[87]
Association test summary statistics (three populations: Kenya, Gambia, Malawi), "A novel locus of resistance to severe Malaria in a region of ancient balancing selection", Nature (2015)	5130 cases, 5291 controls	EGAD00010001081	[88]
Imputation-based meta-analysis of severe Malaria (Gambia) & 1247 cases	1533 controls	EGAD00010000572	[89]
Imputation-based meta-analysis of severe Malaria (Kenya)	1711 cases & 1544 controls	EGAD00010000570	[89]
Gambia Case Control Study	1059 cases and 1496 controls	EGAD00000000087	[86]
Gambia Trios	658 trios	EGAD00000000019	[90]
Ghana Trios	608 trios	EGAD00000000020	[90]
1000 genomes Project	2,504 individuals from 26 populations		[91]

b. Identifying selective pressure variants from literature and databases to construct malaria specific subnetwork of the pathogen:

Selective pressure variants in *Plasmodium falciparum* will be retrieved from literature and essential genome databases provided in **Table 4** genome database category. The single nucleotide polymorphisms (SNPs) detected to show causality from the GWAS data will

be annotated, interpreted and assessed using bioinformatics tools and other tests described by Chimusa et al. [92] to construct gene network.

c. Retrieving and integrating inter and intra functional protein-protein interaction data between host and pathogen into a unified framework:
Reviewed protein sequences of human and the parasite will be retrieved from Uniprot database [93]. La Count et al. [94] experimental *Plasmodium falciparum* protein-protein interaction (PPI) data, Bossi and Lenner human experimentally derived PPI data [95] together with other inter-specie and intra-specie PPIs from REACTOME database [96] and other essential functional databases shown in **Table 4** protein-protein interaction database category will be used. These functional datasets will be integrated using a large-scale data-driven computational framework described by Mazandu et al [97].

Table 4. Useful resources and their descriptions required for this research.

Category	Resource	Description	Reference
Protein sequence database/Protein family and domain database	UniProt	Centralized resource for protein sequences and functional information	[93]
	Interpro	An integrative protein signature database	[98]
Functional genomics database and tools	Gene Ontology database	A classification system for annotation of genes and gene products with molecular function, biological process and cellular component	[99]
	KEGG Database	An integrated database of genes and metabolic pathway information	[100]
	MetaCyc	A universal database consisting of enzymes and experimentally annotated metabolic pathways	[101]
	BioCyc	A database which contains predicted metabolic network of an organism of interest, including metabolic pathways, enzymes, metabolites and reactions	[101]
Protein-protein interaction datasets and database	STRING	Retrieval of functional associations inferred from sequence and high-throughput data	[102]

Continued on next page

Table 4 – continued from previous page

Category	Resource	Description	Reference
	Wuchty experimental PPI	Experimentally derived *Plasmodium falciparum* PPIs derived using protein domains	[103]
	Wuchty et al *in silico* PPI	Computationally derived *Plasmodium falciparum* PPIs derived using protein domains, interologs and experimental PPIs	[104]
	Wuchty et al experimental PPI	Experimentally derived *Plasmodium falciparum* PPIs	[105]
	Wuchty et al experimental PPI	Experimentally derived *Plasmodium falciparum* PPIs	[106]
	Bossi and Lenner human experimental PPIs	Experimentally derived human PPIs	[95]
	Lacount Protein interaction network of *Plasmodium falciparum*	Experimentally derived PPIs	[94]
	mRNA expression profile	Experimentally derived asexual stage microarray data	[107]
	P. falciparum sexual development transcriptome	Experimentally derived sexual stage microarray data	[108]
	REACTOME	Database of manually curated, peer-reviewed pathway database of human pathways and processes	[96]
	IntAct	Protein interaction database system and analysis tools for molecular interaction data	[109]
	MINT	Molecular interaction database of experimentally verified PPIs mined from scientific literatures	[110]
	BioGRID	Curated biological database of PPIs, genetic interactions, chemical interactions and post-translational modifications	[111]

Continued on next page

Table 4 – continued from previous page

Category	Resource	Description	Reference
Drug target database	Tropical Disease Research targets (TDR)	Database of diverse datasets to facilitate the identification and prioritization of drugs and drug targets	[112]
	Drug Bank	Resource combining detailed drug data with comprehensive drug target information	[111]

d Constructing overlapping network between the host and pathogen:
An overlapping network will be constructed using python scripts and the large-scale data-driven computational framework.

Specific Aim 2: Identify protein targets from the overlapping functional network of host and pathogen (leveraging disease-target-drug relationships)

a Integrating all generated subnetwork to identify densely connected subnetworks:
All interaction networks generated in previous steps will be unified using custom python scripts and the computational framework characterized by its ability to integrate heterogeneous data.

b Elucidating clusters or densely connected subnetworks containing important functional proteins by computing functional similarity scores:
The unified network will be partitioned into different functional networks using Blonde et al [113] clustering method to detect connected subnetworks. This step will help identify complex properties including robustness, adaptability and regulation most often observed in living systems.

c Disease-associated gene annotation and pathway enrichment analysis:
Essential biological processes involving the extracted proteins will be predicted using gene functional genomics database shown in **Table 4** under functional genomics database category. Pathways enrichment analysis will be performed on the extracted essential functional proteins to identify statistically relevant pathways which are most likely participate in malaria pathogenesis.

Specific Aim 3: Predict potential drug candidates for identified targets

a Predicting repurposable approved drug candidates by leveraging drug-target-disease associations and pharmaceutical datasets:
The computational framework would be used to map approved drugs to the set of *Plasmodium falciparum* proteins identified implementing semantic similarity method.

1.10 Summary

In this chapter, we described the background of malaria, the origin and spread of resistant variants into Africa, malaria drug resistance and some malaria-specific variants associated with drug resistance and susceptibility. We also described the rationale for this study, the project pipeline and the various tools and source of datasets implemented.

CHAPTER 2

2 Reviewing computational/ *in silico* methods of drug discovery

2.1 Introduction

Drug research and development pipeline entails the following steps: (a) target identification and validation, (b) hit to lead molecule generation, (c) lead molecule optimization and characterization, (d) drug formulation and delivery, (e) pharmacokinetics and drug disposition, (f) preclinical drug candidate identification, and (g) bioanalytical testing and clinical trials [114]. Computational drug discovery has over the past few decades become very relevant mainly due to the reduced risks, time, and resources as compared to the traditional experimental approaches [115]. This has been made possible due to the improved computational power and *in silico* methods.
These steps will complement experimental approaches by streamlining the research scope and guiding *in vivo* validation [116].
Discovery of sildenafil and thalidome are some of the successes in the application of computational approaches to drug design [117]. Traditional novel drug development from scratch to its availability in the market costs about $2.558 billion over a period of 10 to 15 years [114]. With this large investment, the success rate of a drug progressing to the market is about 13% [114].

Rejection of potential drugs particularly during safety and efficacy assessment in phase II and phase III clinical trial development is associated with unexpected clinical side effects and cross-reactivity. This result is significantly increased attrition rate [118]. These unexpected effects centralize on the drug target which may be disease candidate proteins or genes, biological pathways, disease-associated microRNAs, biomarkers, crucial nodes of biological network or molecular functions [119]. This could be linked to inadequate knowledge on the drug targets, undesirable pharmacokinetic expressions upon target interaction or off target effects. This challenge relies on the methods and population data used to identify targets especially for polygenic diseases and this therefore serves as a major bottleneck in drug development. It is also due to the first fundamental stage of drug development which is identifying and validating drug targets of interest for downstream analysis. This highlights the need for modulating drug targets to improve the disease state observed and achieve the desired biological response by elucidating off-targets as observed in promiscuous kinase inhibitors [120].
Experimental drug target identification approaches rely on the characterization of proteins of interest followed by the experimental validation using techniques, such as gene knock-outs, animal studies and site-directed mutagenesis [121]. However, identifying drug targets through these methods are difficult [97].

In the post-genomic era where there is an exponential increase in open access of biological data generated by bioinformatics pipelines, the drug discovery field has been revolutionized with diverse biological datasets which enables scientists to understand comprehensively the biological system relevant to the disease in focus. Thus, the need arose to implement *in silico* methods that would facilitate designing, redesigning and repositioning of drug-like molecules exhibiting desired bioactivity profiles as well as predicting and validating drug targets [122]. This is particularly critical given the increased incidence of widespread drug resistant strains threatening the efficacy of common drugs. Computational methods have transformed rational and systematic approaches for exploring efficiently the space of drug combinations in combinatorial drug discovery.

In silico methods have led to the repositioning of old drugs [123, 124] as well as the prediction of side effects [125] and anatomical therapeutic indicators of approved drugs [126]. This implies that the inception of computational approaches have contributed immensely to a systematic rational guidance

of the processes and to reduction of the period required for drug's availability in the market [124, 127]. This is possible based on the hypothesis that drug side effects would be minimized if the drug candidate is potent and highly selective [127].

The baseline criteria for selecting drug targets requires the potential target(s) to be essential and indispensable to disease outcome. For instance, in genetic diseases, gene therapy involves identifying genetic variants associated with diseases. However, infectious diseases, require an understanding of the complex interplay between the host and the disease causing organism or pathogen [128]. Host target(s) therefore must be unique and homologous to the microbe. Pathogens protein targets that are homologous to the host are eliminated in the computational drug discovery process to primarily avoid any adverse drug reaction. Additionally, the effectiveness of a drug is highly dependent on the target protein(s) in the microbe or essential biological pathway(s) or process that is key to the survival and propagation of the pathogen in the host system.

Leveraging analytical platforms and omics databases containing biological information, computational approaches have become core components in drug discovery pipeline [121]. For instance, analytical platforms help to elucidate essential chemogenomic relationships between available targets data and potential drug candidates or molecules, thereby facilitating the prospects of identifying novel druggable targets, possible off-targets, drug leads and potential repurposable drug candidates. So, it is expected that powerful computational models including but not limited to network-based and machine-learning methods, would lead to better prediction and understanding of drug-target interactions and underlying disease molecular mechanisms.

Computational approaches to drug discovery have helped to translate biological data into functional knowledge treatment interventions against diseases at a faster rate. This approach is characterized by providing a system view of the disease in relation to the biological system of interest. This helps to elucidate important processes, molecular and cellular networks usually difficult to explore experimentally. The ability to reveal such patterns helps to design predictive models to identify disease biomarkers and potential drug targets [129]. Considering complex diseases which are distinguished by their ability to dysregulate biological functions and pathways, computational methods provide the means to understand the regulatory mechanisms through gene regulatory network analysis [129]. Also, the development of computational integrative framework using biological processes, functional data sets (protein-protein interactions between disease causing pathogens and host) together with pharmaceutical data sets facilitates the extraction of drug targets and the identification of drugs possible for repositioning or repurposing against an infection [97, 121].

In the field of pharmacogenomics and pharmacomicrobiomics, computational techniques have facilitated the prediction of drug metabolism by elucidating inhibitors and substrates of specific enzymes involved in metabolism. This has led to an in-depth understanding of *in silico* evaluation of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties though interactive optimization of leads, therefore mitigating the tendency of drug failure [130].

2.2 Current computational approaches for drug target and potential drug candidate identification

2.2.1 Network based analysis approach

The study of disease mechanisms to develop drugs or vaccines have evolved from single gene or protein analysis to an entire multi-scale analysis of genomics, pharmacogenomics, metabolomics and proteomics relevant to the disease of interest. This approach consists of integrating these different large-scale datasets from heterogeneous sources to generate disease-specific networks, fostering a whole genome-based integrative approach to achieve global view. This disease-specific network, which is a biological entity composed of sub-units connected as a whole, is used to elucidate essential nodes which could serve as targets due to their influence within the network [131].
A typical example is observed in the case of drugs, such as artemisinin combination therapies (ACTs) and clozapine for treating malaria and schizophrenia, respectively, which interact with multiple targets to deliver the required therapeutic response [132, 133]. This integrative approach presents a multi-view perspective of elucidating causal genes, relevant pathways and novel drug targets to overcome drug resistance. Also, it increases the reliability in predicting novel drugs and/or putative drugs as well as engineering drug targets to overcome drug resistance [134, 135].
Integrating different biological datasets requires developing algorithms and systems biology tools together with the use of network analysis and functional genomic databases, as shown in functional genomics database and tools category of **Table 4**, to unify the dataset [135]. These tools (**Table 4**) are used to interpret the interactions within the network by identifying sub-networks and regions of similarity and dissimilarity that best explains the disease of interest, to narrow down the research scope for further enrichment and validation analysis to improve disease classification, disease-associated gene prioritization and drug discovery [135].

Network based approach is recommended when identifying targets and drug candidates for most complex diseases [135]. It allows to uncover biological mechanisms involved in development and differentiation of complex diseases [129] The technique is implemented in analyzing nodes and edges in various types of networks including chemical structure and reaction networks, protein structure networks, protein-protein interaction networks, signal transduction networks, genetics interaction networks and metabolic networks.
Moreover, network based approaches sometimes involves computational analysis of metabolisms during the life cycle of the pathogen. Network construction categorizes various metabolic processes into pathways and their reactions and enzymes [136]. This break–down enables analysis of the entire network more conveniently. Flux balance analysis together with *in silico* knock-out studies are implemented in studies during network analysis to identify vital reactions or biological processes essential for the pathogen's survival, thus narrowing down the drug target search space [137]. There are evidences on the use of cellular networks to elucidate complex genotype-to-phenotype relationships among diseases and their associated genetics variants [138]. This technique has become an effective tool for predicting drug-target associations.

Network-based approaches have been widely used to predict candidate targets and drug target interactions. Luo et al. [134] developed an integrative pipeline capable of integrating various data types as well as coping with the noise, incomplete and high-dimensional nature of data sets by learning low-dimensional vector representations of essential features. They identified novel interactions between three drugs and cyclooxygenase of which was experimentally verified and further showed to be potential for preventing inflammatory diseases. Also, various biological network pipelines and algorithms have been developed to predict essential molecular processes and pathways to enhance drug research, thus controlling pathways cross-talk and possible drug resistance [129, 136].

Overall, network based approaches requires a comprehensive understanding of the interaction network particularly regions where potential drug target are located. This, therefore, requires pathway and enrichment analysis to accurately classify the potential drug target. **Figure 4** describes the

summarized workflow of network-based approach.

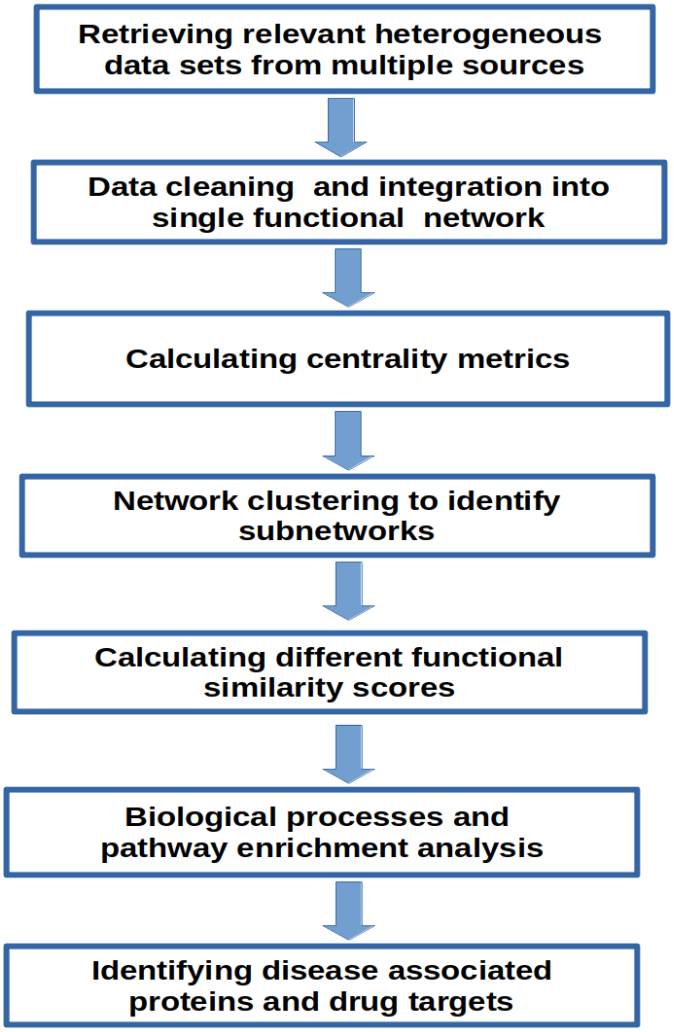

Figure 4. Generalized workflow of network-based approach in predicting potential drug targets and drug candidates.

2.2.2 Data mining (DM)/Machine learning (ML)

With the exponential increase in biological data from high throughput and combinatorial synthesis, the technological and paradigm shift to data mining and machine learning-based methods have enhanced the extraction and processing of these datasets by combining both biological knowledge, computational tools and algorithms. These indispensable techniques are gaining most attention and credibility because of the reliability and accuracy in predicting key property values of compounds and its significant success rate [139]. This is attributed to their abilities to identify and map relationships between large number of compounds which is difficult to obtain using substructural similarities only [139]. Also, machine learning techniques are implemented in both system and molecular methods to predict drug targets through proteomic, microarray and chemogenomic data mining and analysis[119]. In addition, ML approaches have played significant role in the pharmaceutical industry due to essential predictive models, optimization tools and compounds libraries developed to facilitate drug research.

Data mining approaches are primarily characterized by an automatic subsetting of essential information from a pool of datasets. Data mining models ranging from simple parametric equations derived from linear methods to complex models derived from non-linear methods [140] play a critical role in uncovering significant patterns in chemical and pharmacological property space essential

for drug discovery. In addition to that, advanced machine learning models and algorithms such as support vector machines on databases [141], neural networks [139], logistic regression [142], naive Bayesian classification [139, 143], binary kernel discrimination[142], partial least squares [144] and random forest [139] as described in **Table 5** have been significantly instrumental in drug research. For instance, they have contributed to determining pattern recognition underlying the relationship between compounds and calculated molecular descriptors or experimental measurements within large chemogenomic space [122, 140]. ML and DM attempts to find correlations between specific activities or classifications for a set of compounds and their features thus, enabling clustering similarities among drug-like compounds in multidimensional space [122, 140].

For example, Fatumo et al. [145] in their research to identify *Plasmodium falciparum* drug targets developed a machine learning-based metabolic network analysis approach that identified essential reactions/enzymes as drug targets from the metabolic network of the pathogen. The authors identified 46 essential reactions of which 19 had been reported in literature. A study conducted by Sturm el al. [146] applied neural networks machine learning approach to develop an algorithm for microRNA target prediction. The algorithm developed has the ability to predict potential targets sites with or without the presence of a seed match. The model was based on machine learning and automatic feature selection using a wide spectrum of compositional, structural, and base pairing features covering current biological knowledge.

In relation to both structure-based and ligand-based virtual screening, combination of DM approaches and a collection of selective pharmacological agents enables mapping of such chemogenomic libraries into biological activity space to predict potential targets [147]. Particularly, when training sets are available, ML methods are more effective in predicting the physical, chemical and biological properties of small molecules as compared to *ab initio* methods [148]. DM and ML methods are used to develop Quantitative Structure–Activity Relationship (QSAR) or quantitative models for drug-like property predictions and chemical risk assessment [149]. Also, *in silico in vitro* absorption, distribution, metabolism, excretion and toxicity (ADMET) models and *in vivo* pharmacokinetic models for optimizing molecular properties and predicting pharmacokinetic parameters have been developed using ML and DM techniques [150]. These models facilitate the selection of leads with improved strong binding affinity to targets.

Application of DM in target similarity search enables the identification of putative protein targets. This approach involves data mining of pathogen's sequence and querying against drug target databases to identify putative drug targets with suitable druggability index [131]. In a study conducted by Mogire et al. [124] to identify putative drug targets against *Plasmodium falciparum*, target similarity search of the parasites proteome against drug target databases was performed.

ML models are implemented in predicting sensitivity of drug candidates based on cell lines response or the chemical properties of the drugs or a combination of both approaches. This improves the power of designing and systematically analyzing experimental screenings against panels of cell lines in order to identify potential drugs or repurposable drugs [151]. This approach is critical in the area of personalized medicine in terms of leveraging genomic traits to drug sensitivity. Menden and colleagues developed machine learning models that integrates chemical properties of drugs and genomic alterations such as copy number variant and sequence variant from cancer cell lines [151]. Their model predicts sensitivity of genomically characterized cancer cell lines to the drugs in order to ascertain the drug's efficacy [151]. This model have the ability to optimize experimental design of drug-cell screenings by estimating missing half maximal inhibitory concentration (IC_{50}) values [151] In addition, their model predicts essential target-specific association information between compounds and target.

Nidhi and colleagues developed a multiple-category Laplacian-modified Bayesian model that works on the basis of chemical structures to predict targets for all MDDR (MDL Drug Database Report) database compounds [147]. The model generated was trained on extended-connectivity fingerprints of compounds from 964 target families characterized by various levels of annotation in the

WOMBAT (World Of Molecular BioAcTivity) chemogenomics database. It was then used to predict top three most likely targets for all MDDT database compounds. Nigsh et al. [152] compared the predictive power of multiple-category Laplacian-modified Bayesian model and Winnow algorithm, a linear threshold learning algorithm. The Winnow algorithm implements additive machine learning rule in order to minimize ligand-target prediction related errors [152]. It was observed that, both algorithms predict slightly different targets due to compounds that are exclusively retrieved by each algorithm.

Recently, Polypharmacology Browser (PPB2), a new target predicting tool has been reported [153]. This tool implements neural networks and Naive Bayesian classification models to classify ligands based on their molecular fingerprints or descriptors [153]. **Figure 5** describes a summary of application of data mining and machine learning approaches in drug discovery.

Table 5. Summary of data mining/machine learning methods.

Method	Description	Reference
Neural networks and deep learning	A model-based learning method capable of deriving meaning from complicated data based on layers of connected neurons to extract patterns and detect trends that are complex to easily observe.	[139]
Logistic regression	A statistical method with a binomial response variable. This method facilitates easy handling of two explanatory variables simultaneously	[154]
Naive Bayesian classification	A probabilistic classifier that makes classification using Bayesian theorem with strong independent assumptions between the features	[139, 143]
Binary kernel discrimination	Used to generate models that can be used to predict the likely activity of compounds based on the calculation of similarities	[142]
Principal component analysis	A statistical procedure that implements an orthogonal transformation to transform a set of possibly correlated variables into uncorrelated variables (principal components)	[139]
Hierarchical cluster analysis	A classification technique that group data points into clusters either by the agglomerative hierarchical clustering technique or the divisive hierarchical clustering technique	[139]
Partial least squares	A statistical method for constructing predictive models	[144]
Random forest	Learning method for classification and regression based on decision trees	[139]
k-nearest neighbor	A non parametric technique for classification and regression where an object is classified by rules among it nearest neighbors(k)	[139]

Continued on next page

Table 5 – continued from previous page

Method	Description	Reference
Support vector machines on databases	A supervised machine learning algorithm which can be used for both classification or regression by finding the hyper-plane that differentiate the two classes	[141]
Unsupervised learning k-means clustering	A classification method that group similar data points according to a fixed number of clusters (k)	[139]

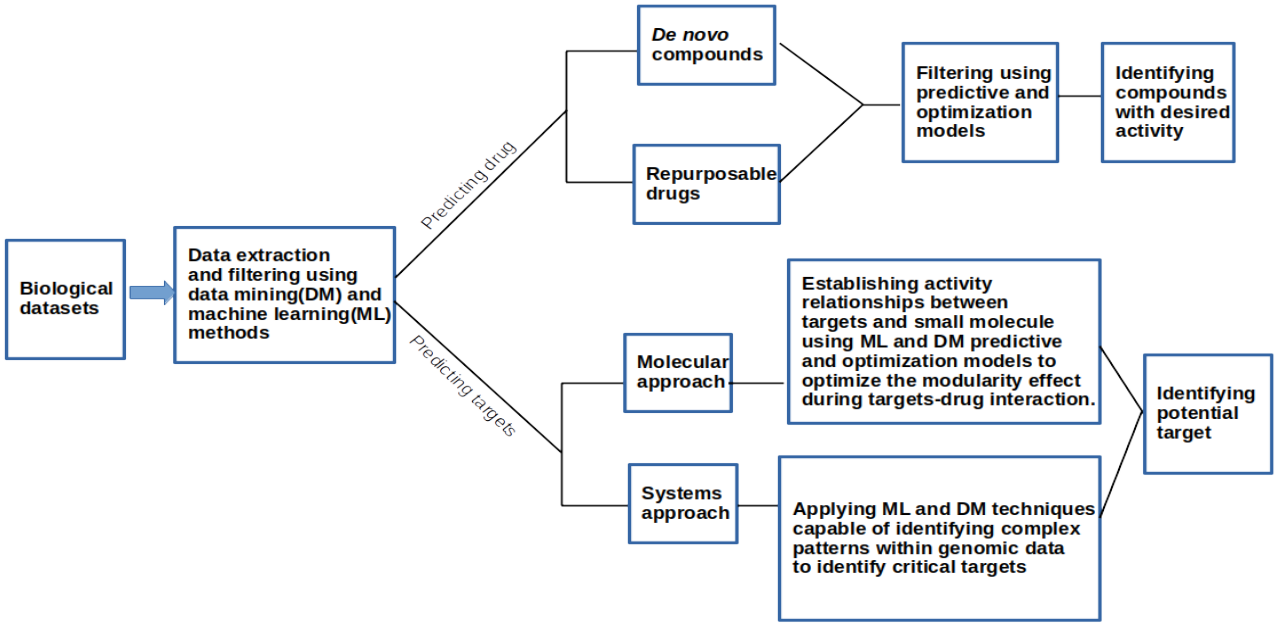

Figure 5. Generalized workflow of data mining and machine learning methods to biological data in predicting potential drug targets and drug candidates.

2.2.3 Reverse / Inverse docking

This computational approach is used for identifying putative binding proteins from protein or genomic databases for particular small molecules with known biological activity [155]. Reverse or inverse virtual screening or inverse docking is a technique that facilitates developing hypothetical relationships among protein targets by chemical probing [156]. It is the structural-based approach of virtual screening unlike the ligand-based methods which require pharmacophores, two dimensional (2D) fingerprints and three dimensional (3D) similarity search [117]. It aims to identify drug targets by screening drug-like molecules against rightful protein databases [157]. Molecular docking simulation involves an optimization process of finding the most favorable 3D binding conformations of the ligand to the target [150]. The targets are assessed and scored using scoring functions algorithms and ranked according to best binding modes and interaction [158]. Interestingly, reverse docking outputs could be used as a profile to characterize the druggability index or enzyme promiscuity of the protein

structures [159]. Reverse docking approach remains a valuable computational technique for exploring alternative uses for existing drugs in terms of drug repurposing and drug rescue [117]. It therefore plays a vital role in the discovery of novel drugs, drug leads, natural products, and other ligands for treating neglected diseases which most pharmaceutical industries are hesitant to invest in due to the fear of inadequate return–on–investment [117].

Unlike the conventional forward docking approaches wherein variety of ligands are docked to a target, reverse docking process involves screening to a set of different protein targets a ligand or compound to identify potential partners through statistical analysis of binding modes within the targets [155]. The framework of this technique is dependent on the knowledge of the distribution of non-homogeneous proteins, their complexities due to the combination of domains and their conformational flexibility due to multiple folds [160]. Due to that, a ligand fits or docks into the functional binding pocket of a specific protein fold based on its three-dimensional conformation and thus, interacts with specific protein residues [156]. This technique enhances elucidation of mechanisms of action and control possible off-target effects. Various tools described in **Table 6** including TarFisDock [157], IdTarget [161], INVDock [162] and AutoDock [163] have been proven to be very useful in drug leads and target prediction [164, 165]. These tools implement different scoring functions to approximate the standard chemical potentials of the system [163].

For instance, idTarget implements the AutoDock4 robust scoring functions [161, 166]. These functions have been shown to have better statistical performance in terms of binding mode prediction [166] and binding site searching efficiency even at the dimensionality of 30 [167].

TarFisDOCK, a valuable tool for target prediction was developed from DOCK version 4. The tool however is still under improvement due to associated false positives as a result of inaccuracies in the scoring function for reverse docking [157]. These errors could be associated with less coverage due to limited target datasets and inability to incorporate protein flexibility during docking.

INVDOCK implements a scoring scheme capable of performing binding competitive analysis as well as evaluate the interaction energy between docked structures [162]. It is based on the concept of binding competitiveness such that, a drug that binds to its target non-competitively is likely to be less effective.

AutoDOCK implements a machine learning based scoring function that explores the iterated local search global optimizer approach [163]. The scoring scheme is based on the advantages of knowledge-based potentials as well as extracting empirical information from conformations of receptor-ligand complexes and the experimental affinity measurements [163].

Inverse docking has the ability to predict off-targets for ligands aside facilitating predicting activity and selectivity of unknown ligands against known targets [168]. It has been applied in evaluating the binding energies (usually expressed in kcal/mol) and modes of libraries of compounds against panel of proteins. This evaluation results in a defined group of protein-ligand complexes thus, enhancing the identification of lead compounds for subsequent biological test. This application reduces cost involved in compound development and biological screening as well as reduces synthetic efforts and time required for *de novo* drug discovery [156]. Lauro et al. [168, 169] applied this method to natural bioactive molecules to investigate their efficacy against a panel of cancer associated proteins. The idea of polypharmacology led to the development of selective optimization of side activities (SOSA) approach which enhances the generation of new biological activities [170]. The challenging part of this approach is the construction of panel of target proteins taken in to account careful selection of proteins not belonging to the same folds. Also the accuracy and reliability of this approach is limited when 3D structures of the protein targets are not available.

Regardless of the advantages of reverse docking methods in *in silico* drug research, they are complex as compared to forward docking techniques such that larger target structure datasets are required to increase the coverage and predictive power [155, 159]. Also, aside its associated biases in inter-protein scoring yielding false positives [171], it requires high computational cost [159, 172].

Table 6. Description of various tools applied in reverse/inverse docking.

Tool	Description	Reference
TarFisDock	A web-based tool that offers potential drug target databases together with a reverse ligand-protein docking approach which are used in searching for small molecule-target interactions. The tool accept small molecule as input and predicts possible binding drug targets.	[157]
IdTarget	A web-based tool for predicting potential binding proteins for chemical compounds through a divide-and-conquer docking approach and robust scoring function based on regression analysis and quantum charge models. It has the ability of screening against majority protein structures in Protein Data Bank	[161]
INVDock	This software enables automatic identification of potential protein, RNA and DNA targets of small molecules by searching in both protein and nucleic acid 3-D databases. It implements a flexible docking algorithm	[162].
AutoDock	A suite of automated docking tools to predict binding interactions between small molecules and targets of known 3D structure	[163]

2.2.4 Biological activity spectra (Biospectra) analysis

There are several reported evidences to the fact that most drugs establish therapeutic response through multiple target modulation [173, 174]. The ability to predict such functional consequences of biological perturbations between the genome or proteome of an organism and biologically profiled compounds is indispensable in drug research. In relations to that, analysis of the modularity effect drug-like molecules impact on target's function is a must to understanding the expressed phenotype or therapeutic response capacity of the molecule [175, 176]. Biospectra simply refers to the activities of compounds across potential targets which could enable investigating structure-property relationships [177]. Biospectra analysis is a probabilistic structure–activity relationship approach that complements experimental affinity-based studies [176]. The technique herein is used for measuring quantitatively the patterns and dynamics of the functional activity of a molecule across multiple potential targets [176, 177]. It is therefore a determinant of the inhibitory or stimulatory effect profiles of drug–like molecules on targets within a system. Studies have shown that the association between proteins, drug-like molecules and biospectra serves as the building blocks for developing probabilistic approaches to drug discovery [178].

This method provides a firm foundation in determining quantitatively the correlation between molecular structures and biological effect profiles by providing estimates of the therapeutic effect of a molecule. Estimation is done by constructing a nonlinear multivariant model which provides an unbiased tool for investigating associations between structure and function similarities of molecules [176]. Such analysis is relevant for predicting drug targets for orphan compounds based on the concept of

chemical structure similarity [177].
It provides the means to classify molecules on the basis of biospectra similarity as well as predict interacting capabilities of molecules with multiple targets. This classification mechanism allows for identification of molecules with similar function with no prior information concerning the target which is difficult using experimental techniques. Biological activity spectra is an essential indicator of molecular property descriptor [179]. This method was implemented by Fliri and colleagues to identify agonist and antagonist effect profiles of medicinal agents on brain dopamine receptors belonging to the *GPCR* superfamily [176]. This technique facilitates the ability to conduct spectra similarity and hierarchical clustering methods through profile similarity measurements thus, establishing quantitative relationships between chemical structures and biological activity spectra [179]. Biospectra analysis have been shown to be critical in mining pharmacology datasets as well as predicting possible adverse drug effects based on profile similarity with drug-like molecules known for adverse reactions [177]. Similarity between molecules are measured using Tanimoto similarity coefficient [180], cosine correlation [181], Euclidean distance [182] or city block distance [183].

Paolini and colleagues presented a comprehensive mapping of pharmacological space by applying probabilistic model on integrated structure-activity relationships data [178]. They found 836 human genes discovered verified targets for small molecules. This integration enables the identification of unique molecular targets through construction of a ligand-target matrix.
Since similar drug-like molecules express similar biospectra, this approach is useful for drug repurposing because it facilitates the translation of biological response data into chemical structure design [176]. This implies that, the ability to correlate off-target effects with biological spectra would help map unto new targets where the response might be beneficial to address a different diseases. For example, sildenafil initially developed to treat angina expressed a side effect of prolonged penile erection and this resulted in a change of the treatment focus of the drug [184].

However, biospectra analysis is highly dependent of experimental data obtained from various ligand-binding assays or a martix of targets which could be difficult.

2.2.5 Ligand-based *in silico* target prediction

Ligand–based computational approach is the framework for ligand-based drug discovery. It is based on the concept of chemical structure similarity, which states that similar ligands or compounds would bind to similar targets with almost the same binding affinity and express similar biological responses [160]. This concept of similarity has been extensively utilized in lead discovery and optimization primarily because it takes into account the polypharmacological nature of drugs [185]. Also, it is essential for quick investigation on primary and secondary targets as well as selectivity among target families [177]. The approach herein involves the interplay between characterized protein targets and characterized ligands with similar chemical structure, properties and pharmacophoric features to enable predict biological targets. This is achieved by mapping the structures of compounds known to modulate cellular phenotypes (mostly natural products or orphan compounds) unto chemogenomics databases containing biologically profiled compounds with known targets [177].

In that regard, cheminformatics and bioinformatics have developed mapping models including but not limited to topological-based models, Bayesian classification models and atom pair-based models from available bioactivity data using machine-learning and statistical methods [173]. These models are implemented in mapping compounds into chemogenomical space or bioactivity database taking into account either 2D or 3D molecular descriptors [160, 186] and chemical fingerprints [187] for measuring similarity among structures to predict targets. An advantage of using chemical fingerprints in designing models is that it enables back-projections of correlation between characterized proteins and compounds unto orphan compounds with the knowledge that similar compound structures would exhibit similar affinity chemical fingerprints [177].
Molecular descriptors are numerical features extracted from the compounds based on their molecular

properties [160, 188] whereas chemical fingerprints are high dimensional vectors that encodes the presence of substructural fragments [187, 189].

Ligand-based approaches to target prediction provides the platform to understand the relationships between structurally dissimilar but functionally related proteins based on their ligand similarity, thus, helping to form hypothesis which can be verified using statistical methods. Similar to biospectra, ligand-based approach is more informative for pharmacology, medicinal chemistry and biochemistry [173].

Similarity among 3D structures are measured using Minkowski distance metrics and Tanimoto similarity coefficient and its complement, the Soergel distance [160]. Tanimoto similarity coefficient can be applied to 3D structures [139] however, this metric is susceptible to molecular size because it fails to account irrelevant features of a large molecule, thus, resulting into odd size dependencies [160].

These measurable features of compounds have been implemented in developing tools such as Similarity Ensemble Approach (SEA) [173], Swiss Target Prediction (STP) [190], SpiDER [191], SuperPred [192], Polypharmacology Browser [193], HitPick [194], Prediction of Activity Spectra for Biologically Active Substance (PASS) [195], MOst-Similar ligand-based Target inference approach (MOST) [196], Candidate Ligand Identification Program (CLIP) [197] and Chemical Similarity Network Analysis Pulldown (CSNAP) [198]. These tools described in **Table 7** implements fingerprints and/or structural similarity to predict ranked targets from ligand-target datasets in order of decreasing similarity score.

Ligand-based target prediction approach is not feasible in the cases of predicting targets with no or only a small number of bioactive ligands and ligands that exhibit activity cliffs characterized by high structural similarity but different activity. [122, 199].

Table 7. Description of various tools applied in ligand-based approach.

Tool	Description	Reference
Similarity Ensemble Approach	This method is based on different types of molecular fingerprints to predict targets	[173]
Swiss Target Prediction	A web-based server for predicting targets for bioactive small molecules. This tool is based on a combination of 2D and 3D similarity measures or features with known ligands	[190]
Self-organizing map–based prediction of drug equivalence relationships(SpiDER)	This computational tool is used for predicting macromolecular targets for both de novo designed molecules and known drugs. It merges concepts of self-organizing maps, consensus scoring and statistical analysis.	[191]

Continued on next page

Table 7 – continued from previous page

Tool	Description	Reference
SuperPred	This tool is based on the hypothesis that similar structures have similar activity profiles. It is a web-based tool that translates molecules into molecular descriptor or fingerprints and then maps unto drugs with established molecular targets links. Similarities between compounds and existing drugs are expressed using Tanimoto coefficient in order to identify potential targets.	[192]
Polypharmacology Browser	A versatile web-based tool that predicts potential drug targets for small molecules by searching nearest neighbors using 10 different fingerprints which integrates the composition, substructures, molecular shapes as well as pharmacophore features.	[193]
HitPick	A web-based tool that implements the B-score method and other statistical methods for identifying hits in high-throughput screenings and predict their potential targets	[194]
Prediction of Activity Spectra for Biologically Active Substance (PASS)	A web-based tool for predicting the biological activity spectra of compounds on the basis of structural formula. These predictions are efficient in finding new targets and/or new compounds for targets.	[195]
MOst-Similar ligand-based Target inference approach (MOST)	A computational tool that uses fingeprint similarity and bioactivity of most-similar ligands to predict targets for query compounds	[196]
Candidate Ligand Identification Program (CLIP)	A target prediction tool that implements the similarity search approach using the Bron-Kerbosch clique detection algorithms to find structures common to that of a known bioactive target 3D structures. This is achieved by characterizing pharmacophore points of structures for searching	[197]
Chemical Similarity Network Analysis Pulldown (CSNAP)	A drug target prediction method that is based on chemical similarity networks for large-scale consensus chemical pattern recognition through clustering used for drug target profiling.	[198]

2.2.6 Target-based *in silico* prediction

In contrast to ligand-based *in silico* prediction, target-based approach involves predicting ligand or compound partners from the targets perspective noting their complexities resulting from combination of multiple domains, promiscuity of fold families and conformational flexibilities [160]. This is because, protein conformational changes coupled to ligand binding constitutes the structural stabiliza-

tion and energetics basis underlying protein regulation [200]. In this approach, targets are predicted by investigating protein functionalization when ligands or substrates bind to a well characterized binding pocket. Due to that, several algorithms such as Bron–Kerbosch clique detection algorithms [197] are used to build binding site similarity methods which are implemented to identify significant targets. These methods including CavBase [201], SuMo [202], IsoMIF [203], PocketMatch [204], Pocket Alignment in Relation to Identification of Substrates (PARIS) [205] captures local physico-chemical correspondence and functional relationships among proteins structures or substructures (especially regions in binding sites) as well as local spatial similarities.
Other techniques implement functional pharmacophore screening approach. This approach entails using pharmacophores of small-molecule drugs to investigate their binding interactions within binding pockets of targets and identify such compounds whose features aligns with the description [160, 206]. PharmMapper [207] tool has been developed to implement such an approach. This approach has been extensively used to search for targets for components of Chinese medicines.
Predictive models have been modified to incorporate functional effect prediction (activation or inhibition) of compounds on the target since this may positively or negatively modulate a pathway which may result in a desired or undesired functional activity [208, 209]. A recent study by Mervin et al. [210] used models ranging from simplistic random forest to cascaded models which implements separate binding and functional effect classification steps to predict functional effects. **Table 8** provides a description of the various tools used for target-based approach.

Table 8. Description of various tools applied in target-based approach.

Tool	Description	Reference
CavBase	A database that stores set of cavities retrieved from Protein Data Bank for detecting functional relationships among proteins. It is based on the hypothesis that protein functions are associated to their ability to recognize small molecules as well as the interactions established in a defined binding pocket. The method queries a protein cavity to extract similar ones using clique detection algorithm.	[201]
SuMo	A web-based tool used for finding arbitrary 3D structures or substructures of proteins. It is based on how macromolecules are uniquely represented using selected triplets of chemical groups.	[202]
IsoMIF	A web-based tool for identifying molecular interaction field similarities. The tool is useful in predicting functions of targets, identifying polypharmacological targets or cross reactivity as well as identifying potential repurposable small molecules.	[203, 211]
PocketMatch	An algorithm that has the ability to infer functional similarities between protein structures in a high-throughput manner by accurately and efficiently performing large-scale comparison of their binding sites	[204]

Continued on next page

Table 8 – continued from previous page

Tool	Description	Reference
Pocket Alignment in Relation to Identification of Substrates (PARIS)	A tool capable of quantifying the relationships between binding pockets to investigate there essentiality for ligand prediction. In this tool, each pocket is represented as cloud of atoms. Similarities are measured by by aligning these atoms in 3D space and analysed using convolution kernel	[205]
PharmMapper	A computational pharmacophore web-based tool for drug target identification. It is based on a reverse pharmacophore principle where by the query compound is mapped against an annotated pharmacophore model database thus, enabling polypharmacology prediction techniques critical for drug repositioning and potential offtarget risk prediction	[207]

2.2.7 Genomic analysis approach

In this post genomic era, high-throughput hybridization-based technologies and deep sequencing methods have led to the generation of large genomic datasets. Systematic analysis of gene expression and transcriptome of an organism is key to identifying relevant pathways in disease pathology and characteristic expression patterns of specific disease–associated genes. This supports developing diagnostic and prognostic biomarkers critical for disease treatment specifically in areas of personalized or precision medication [212]. Also, analysis of these datasets enhances the identification of functional genetic variants (derived and risk alleles) and prediction of traits. For instance, an extensive study on the morphology of cancer has led to the understanding that, it is as a result of multiple genetic anomalies and as such, individuals with the same cancer type have different genetic anomalies in their tumour [213].

Genomic analysis is a data-quality dependent technique that involves both knowledge-based for the case of genes known to be associated to the disease of interest and data-driven unbiased approaches for cases of no prior knowledge of contribution of a gene to the disease of interest. It requires no assumptions of a genes role. Comparative genomics analysis between cases and controls together with advanced computational models and various genome reference panels, provides the platform to interpret and analyze disease-associated gene variants identified through genome wide association studies (GWAS) or next generation sequencing (NGS) or whole exome sequencing (WES) or transcriptomic studies (RNAseq). This helps to translate these functional knowledge into treatment interventions particularly leveraging disease-target associated data sets [128, 212]. Genes identified could be prioritized to identify putative drug targets and vaccine candidate targets [214]. However because GWAS is not able to explicitly identify causal relationships, the combination of different methods would increase the predictive power of identifying highly potential targets. A study conducted by Fan-Minogue and colleagues evaluated the effectiveness of differential gene expressions (DGE) data, disease associated single nucleotide polymorphisms (SNPs) as well as the combination of the DGE and SNPs to predict drug targets for 56 human diseases [215]. Their studies showed that the combination provided higher predictive statistical power for prioritizing candidate targets as compared to individual DGE and SNPs datasets.

Studies have shown that genes with disease associated alleles are highly potential drug targets [216, 217]. Careful analysis of gene expressions data, somatic mutations data as well as genetic associa-

tion data has been widely explored to study genetic causes of disease and also identify drug targets [218, 219]. In view of that, genomic approaches to drug discovery is a delicate field to address genetic diseases particularly, those with few effective therapies such as neurodegenerative diseases. Genome analysis provides the platform to identify genes that encodes novel proteins or regulatory elements encoding potential drug targets.

Comparative genomics method, a well studied analytical approach enables integration of omics data and the use of several bioinformatics tools such as Identity Plot Maker and Visualization Tool for Alignment (VISTA) [220]. This method enables comparison of two or more genomic sequence of an organism to discover the similarities and differences among the genome thus, exploring and understanding the significance of biologically conserved active regions [220]. In addition, it helps examine the broad spectrum and selectivity of potential genes or proteins as targets across various species of an organism. Comparative genomics therefore provides a means for studying evolutionary changes among organisms by investigating derived and risk alleles which helps to identify genes that are conserved or common among species, and also genes that contribute to the uniqueness of an organism.

Subtractive genomics on the other hand is a widely used approach in drug discovery to identify novel drug targets [137]. This *in silico* approach is based on the identification of essential and non-homologous proteins within the pathogen of interest [221]. Various tools required for genomic analysis are described in **Table 3** and **Table 4**.

2.3 Comparing different approaches in computational drug discovery

As described previously, several computational drug discovery approaches have been suggested, including genomic, biospectra, network-based, machine learning / data mining and virtual screening/molecular docking simulation approaches. Although these methods individually have their specific areas in drug discovery that best describes their usefulness, they have the ability to be integrated to understand complex biological system in order to address challenges in computational drug discovery. This is because, technological advancement has led to the generation of various datasets types describing biological systems from different dimensions some of which are sequencing, gene expression activity and proteomics [222].

In the area of predicting and assessing pharmacological effects of a drug, the combination of these techniques has been instrumental in determining drug target interactions (DTIs) with high efficiency and low cost. In comparison to experimental techniques (*in vitro* and *in vivo* methods), computational methods have provided the technicalities to systematically determine all possible interactions in order to clearly elucidate the pharmacological patterns [223]. Higher dimensional levels of prediction revolve around systematic analysis of biological complex networks and large integrated biomedical datasets, and as such, using a combinatorial approach is highly essential. Some approaches share similar concepts but applied in different forms in addressing similar issues, thus, combination helps to compensate for individual limitations. This in turn increases the accuracy of predicting and minimizing possible adverse effects [224]. For instance, molecular docking principles in elucidating DTIs require 3D structures. Due to that, there are associated biases and false positives when high quality 3D structures are not available [223]. Unlike molecular docking approaches, ML drug target interaction predictive models have the extended capacity of taking into consideration not only the 3D structures of targets but also molecular and protein sequence descriptors [223]. However, network-based methods in predicting DTIs to investigate pharmacological effects apply recommendation algorithms implemented in recommender system [225] and link to prediction algorithms [226] rather than 3D structures and molecular systems. Also, network-based methods are relatively faster compared to the other methods. This is because, the DTI network of interest can be represented as a matrix on which calculations can be computed easily [223]. Additionally, they have the extended capacity of predicting drug effects through simple dynamic processes such as random walk, resource diffusion

and collaborative filtering on biological networks [223].
For example, Paolini and colleagues studied polypharmacology interaction network for human proteins by constructing ligand-target matrix using a Laplacian-modified Bayesian probabilistic models to explore the relationships between chemical structure and targets by integrating diverse structure-activity relationships data [178]. They observed 35% of 276,122 active compounds within their database to hit more than one target whiles 65% hit a target, thus indicating extensive promiscuity of drugs and leads across targets.

Data mining is highly essential in chemogenomics to mine chemogenomic datasets. This is critical in establishing the relationship between set of potential drug targets and ligands. However, the interplay among a holistic picture of the biological system (network), molecular docking approaches and ML methods in chemogenomics presents a broader scope to investigate the effects of compounds on gene/protein expression. To efficiently identify and assess the effects of specific protein targets on specific drugs, robust molecular docking systems that implements ML and DM models have been developed to optimize the performance of predicting drug's effect across molecular networks [227]. These models provide the platform to avoid unnecessary assumptions by specifically accounting for binding effects most often challenging to model without ML and DM techniques. Utilization of this approach in designing scoring functions have significantly enhanced the accuracy of establishing binding affinities of various protein-ligand complexes [228]. Ballester and colleagues developed a competitive high performance scoring function that implements random forest to capture binding effects [228]. The flexibility of their scoring function compared to other rigid functions ensured that it has high predictive power when tested on trained datasets. Another application of this approach is developing machine learning-based scoring and binding affinity functions integrated with molecular docking tools to address difficulties involved in molecular docking. Hsin et al. [224] developed a computational screening approach using machine learning and docking packages to investigate the polypharmacological nature of compounds against potential targets within a biological network. The model developed, has the ability to assess binding modes and predict the best binding mode to targets. This approach increases the reliability and confidence in assessing the binding conformations of compounds and predicting best modes [227]. It also helps to rate the performance of various docking packages as well as compensate for scoring functions–associated errors [229, 230]. Advanced ML methods provide the technique to investigate drug effect in preclinical research and clinical trials. Also, they provide an efficient way to systematically and analytically extract meaningful biological information from clinical trial datasets. This facilitates the ability to design the chemical structure of drugs to modulate drug-target interactions. However, it is of importance to that, the ability to interpret such datasets is challenging and as such requires experience and high technical skills.

Deep learning, a class of machine learning, has strong generalization ability and feature extraction capability. It has emerged as a powerful tool capable of identifying highly complex patterns in both homogeneous and heterogeneous datasets. In computational drug discovery, deep learning method has enhanced prediction of bioactivity, *de novo* molecular design, virtual screening, activity scoring and synthesis prediction [141, 231]. This is mainly because it has less generalization errors, thus, yielding impressive results as compared to traditional machine learning. It has been extensively applied in functional genomics in discovering DNA-binding motifs, and determining sequence specificity of DNA and RNA binding proteins [232].

Data mining and machine learning models are implemented in computational drug discovery for unbiased mining and analysis of genetic datasets mostly in the focus of personalized medicine [233]. The aim of personalized medicine is to discover novel drugs and biomarkers for specific patient groups, most suffering from complex disorders. Developments in this field are applied most often in gene and immuno-oncology therapies for highly personalized and specific group treatments respectively [234]. In view of that, genomic approach together with ML methods provides the platform for identifying disease associated genes (particularly rare disease variants) and their corresponding mutations from methods like DNA sequencing and GWAS [234]. This helps to translate functional

results into treatment and strategic measures. Analyzing genetic functional networks with ML and DM methods enhances the chances of identifying novel biomarkers and drug targets. For example, combination of network-based approach and ML methods plays an increasingly significant role to predict novel mechanisms underlying disease-specific targetable genes or pathways associations. This in turn offers the opportunity for finding new applications for drugs as well as predicting potential adverse effects. Bari and colleagues developed a machine learning-assisted network inference algorithm capable of identifying Class II cancer-associated genes in a cancer network generated from support vector machine models [235]. Also, this combination have been implemented in target fishing using chemical fingerprints [236].
Integration of ML and DM approaches together with network-based techniques is of noteworthy importance in analyzing biological networks to identify potential set of genes or pathways that could serve as targets in combinatorial therapy. The rationale behind this combination strategy is not only to overcome resistance and limitations of monotherapy regimens but also, to overcome the complexities of diseases such as malaria, tuberculosis, cancer and HIV [237]. In that regards, predictive models based on ML, DM, network-based and sometimes molecular docking approaches have been developed to investigate the synergistic effects of drug-like molecules on specified targets [238, 239]. These models incorporate heterogeneous datasets such as cell signaling pathway, transcriptomic and pharmacological datasets [239]. The models have the extended ability of providing insights into biological mechanisms underlying the synergistic combination.
Combination of genomics approach with molecular docking simulations are mostly applied in discovering novel ligands or drug-like molecules to treat infectious diseases.

2.4 Source of Drug Failure: Challenges and Opportunities

2.4.1 Incomplete knowledge on the biological mechanisms underlying certain diseases:

A critical draw–back in the success story of drug discovery is associated with poor understanding of the underlying mechanisms behind some diseases such as nervous system disorders, chronic kidney disease, idiopathic pulmonary fibrosis and other complex disorders [240]. Inability to elucidate genetic variants, biomarkers, pathways or proteins involved in the aetiology of such diseases continues to be a challenge to drug research. Due to that, specific targeted drugs or vaccines have not yet been developed. Researchers have shown that indepth knowledge of disease mechanisms and the elucidation of critical biomarkers would contribute significantly to drug development [240]. This could be associated with inadequate specific datasets available to help unravel the mystery behind a disorder. Due to that, there is an intensified scientific research into bridging the gap between disease mechanisms and drug development. Combination of genomics, chemistry and clinical datasets together with advance ML techniques has been promising in exploring potential targets. A typical example is Alzheimer's disease, in which various mechanisms are been identified through extensive research [241].

2.4.2 Drug resistance development:

Drug resistance has been a major burden in drug use. More often, drugs, particularly, those targeting disease causing pathogens in infectious diseases lose potency with time primarily as a result of selective pressures resulting in drug resistant strains development. This challenge contributes to disease resurgence and increased morbidity and mortality rates. In complex diseases, drugs targeting human cells develop resistance through factors like epigenetics, DNA damage repair and epithelial-mesenchymal transition [242]. In general, drug efflux and drug inactivation are common factors linked to drug resistance. This phenomenon continuously necessitates further research and alterna-

tive treatment development. In addition, researchers are investigating the core biological associated activities resulting into resistance to identify novel approaches to counter such effects.

2.4.3 Inability to reproduce generated disease-related datasets:

Data reproducibility crisis remains a critical challenge in this post-genomic era. Data validation is a measure of the confidence and integrity of the datasets. It is noteworthy that, inconsistencies in results obtained from replicating experiments in different laboratories breeds unsuccessful translation of discovery research as a result of the level of mistrust in the data [240]. This situation significantly slows the rate of translating biological data into functional knowledge and treatment interventions. However, it is argued out that such differences in results could be attributed to confidence interval defined for the independent study as well as inadequate knowledge in essential statistical methods and tools used. Researchers have proposed that external validation and explicit reporting of experimental datasets could possibly increase reproducibility [240]. Also, this challenge presents the opportunity for researchers to develop standardized procedures tailored to each working environment to ensure reproducibility of results and continuity of scientific knowledge.

2.4.4 Complex unpredicted metabolism networks

Unpredicted interactions and mechanisms within a network due to associated kinetic interactions results in an incomplete picture of the cellular behaviour [243]. However, over assumptions in modelling hinders the ability to develop accurate models to answer the biological hypothesis. As a result, algorithms developed for such models produces results that deviates from the true expectations. This therefore presents a challenge in modelling the system to overcome unknown associated metabolic fluxes. As such, there is a higher likelihood of missing essential informations such as pathway and biological activities essential for drug research. There are off-target metabolic interactions that occur as result of metabolic pathways that lead to modelling challenges. Off-target metabolic interactions can be responsible for expected and unexpected responses which most of the time are side effects. In overcoming these challenges and to minimize drug failures and associated adverse effects, predictive models for individuals target networks that simultaneously detect metabolic similarity of associated metabolic pathways using joint learning algorithms.

2.5 Summary

In this chapter, we presented various computational approaches and tools essential for *in silico* extraction of drug targets, predicting potential drug-like candidates, analyzing bioactivity profile and elucidating possible off-target effects in drug discovery. These approaches complement experimental techniques in drug development.

Furthermore, we highlighted on the application of machine learning, data mining, genomics and network analysis techniques in investigating the dynamic patterns within integrated datasets from multiple sources to predict critical nodes, pathways and biological processes. These techniques are relevant in achieving a global perspective of the biological systems to investigate the interplay between multiple independent genes or proteins on disease aetiology. This therefore provides the platform to elucidate set of functional biological entities for drug and vaccine development.

We discussed various molecular docking simulation techniques. We showed the specificity of each approach in terms of predicting potential drug-like molecules and protein targets in drug development. We emphasized on the application of these methods in drug repurposing and reuse particularly in addressing drug resistance and drug development for orphan diseases thus, contributing to limiting the risk of drug failure during trials. Also, we highlighted the combination of machine learning and molecular docking techniques in designing various predictive models to investigate the structural and chemical properties of ligands or drug molecules and validate their efficacy in drug development.

We have shown that these approaches can be combined to compensate for limitations of individual methods thereby increasing the predictive power.
Finally, we presented sources of drug failure looking at the challenges and opportunities involved.

CHAPTER 3

3 *Plasmodium falciparum* Proteome Functional Networks

3.1 Introduction

Understanding of disease causing organisms to elucidate the mechanisms behind the evolution and emergence of resistant strains in order to develop better candidate drugs involves an extensive knowledge of its biological processes at the cellular and molecular level [14]. This activity requires an in-depth understanding of the organism's proteome which is regarded to execute the genetic programme. However, proteins rarely act alone but in extended networks. They establish complex physicochemical dynamic connections in order to facilitate structural and functional organization of the organism. These connections makes up the protein–protein interaction network (PPIN). The interactome network provides a general review of possible interactions than can occur between proteins [95]. Proteins interact both directly and indirectly within a system to maintain the stability and robustness of the system and as such, they are fundamental and critical to every process in the cell. For our study, we will focus on functional protein-protein networks to ensure potential functional interconnectivity of host and pathogen.

In this chapter, a computational integration technique is applied to construct *Plasmodium falciparum* functional protein network in order to build human-*Plasmodium falciparum* protein network. Primary data sets, such as protein sequences, functional data sets from high–throughput experiments, protein signatures from databases such as InterPro [98] and *in silico* generated functional datasets were implemented to construct a unified functional association protein-protein interaction network.

3.2 Assembling Functional Interaction Datasets for Constructing *P. falciparum* Functional Network

Various heterogeneous parasite datasets from different sources such as literature, databases and high-throughput experiments described in **Table 4** were implemented in this step. The datasets are categorized into functional interaction and genomic datasets as shown in **Table 4**. Parasite functional datasets for this research are those retrieved from PPI databases, functional genomics databases and high-throughput experiments (see **Table 4**). On the other hand, genomic datasets comprises of protein sequence and protein family and domain data retrieved from sequence databases such as Uniprot and Interpro as shown in **Table 4**. We integrated various datasets in this study with the overall aim to increase the coverage, sensitivity and accuracy of the generated network.
163 reviewed *Plasmodium falciparum* isolate *3D7* protein sequences retrieved from Uniprot database were used to generate pairwise sequence similarity associations using Basic Local Alignment Search Tool (BLAST) specifically blastp, the protein-protein BLAST algorithm [244]. Only reviewed protein sequences were considered in this study because they have been manually annotated using literature records and curator-evaluated computational analysis [93]. The pair-wise relationship was based on the hypothesis that, proteins sharing conserved domains or families have higher probability of establishing potential functional associations [245].

Parasite interaction dataset consisting of only reviewed proteins were also retrieved from STRING database version 11.0 [102], for this study upon querying the sequence data. STRING database makes predictions based on known interactions from curated databases, experimental findings, gene neighbourhood, gene fusion and gene co-occurrence. It also makes predictions through text mining, co-expression as well as protein homology. STRING predicted interaction comprised of 386 interactions among 114 proteins. The score for these functional interactions ranged between 0.4 and 0.99.
In addition, the *Plasmodium falciparum* interaction dataset from IntAct database [109] was retrieved for this study. IntAct provides interaction data that are derived from literature curation or direct user

submission. The database is linked to other functional databases such as BioGrid, MINT and Uniprot described in **Table 4**. The retrieved interaction dataset comprised of 2,916 interactions between 1,343 proteins.
Also in our analysis, we included experimentally determined *Plasmodium falciparum* protein-protein interaction dataset described by La Count et al. [94]. This dataset was identified through literature review. This dataset covers about 25% of the parasites proteins. This functional dataset is made up of 2,846 interactions from 32,000 yeast two-hybrid screens with *Plasmodium falciparum* protein fragments. These proteins are known to be involved in the parasite's intraerythrocytic cycle [94].
In addition, a comprehensive *Plasmodium falciparum* protein interaction map predicted by Wuchty et al. [105] was used in our study. The map augments protein interaction information that has been retained by the evolutionary divergent model organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster, Caenorhabditis elegans*, *Escherichia coli* together with experimental *P. falciparum* PPI data. The network consists of 4,918 interactions among 1,872 proteins.
Also, our research study incorporated *P. falciparum* protein interaction datasets generated using weighted protein domains of the Protein Family database (PFAM) [103]. The network is made up of 1,428 protein interactions among 361 proteins. In addition to that, we used another interaction data generated by Wuchty et al. [104]. This data comprising of 19,979 interactions among 2,321 was derived by inferring interologs, protein domain and experimental interactions.
P. falciparum experimental interaction dataset used by Wuchty et al. [106] in a study to identify conserved PPIs in *S. cerevisiae, C. elegans* and *D. melanogaster* that have orthologs in *P. falciparum* through comparative topological analysis was also used for our study.
Parasite interaction dataset was also retrieved from InterPro version 73.0 database [98]. InterPro is a protein signature database comprising of protein families, domains and functional sites from other databases including but not limited to PANTHER and PROSITE described in **Table 4**. The generated functional dataset comprised of 1,013 interaction between 256 proteins. The prediction is based on the idea that each interacting protein pair in the dataset share common domains.
Overall, we were able to ensemble a map of *Plasmodium falciparum* interaction datasets from nine independent studies.

3.3 Scoring Functional Interaction Datasets

Protein-protein interaction (PPI) maps are experimentally determined on a large or small scale using either the binary approach or the co-complex approach [246], the two main technologies for PPI determination. Several experimental methods such as yeast two-hybrid screen approach, RNA expression profiles, genetic interaction and mass spectrometric identification methods implement such approaches to investigate functional PPIs [14, 246]. Yeast two-hybrid and mass spectrometric techniques seek to detect physical binding among proteins whereas RNA expression profiles and genetic interaction techniques aim to detect functional associations between proteins which most of the time take the form of physical binding [247]. Aside these methods being labour and resource intensive, they are limited by their interaction classification biases, noise, sensitivity, coverage, complementarities and accuracy of the data generated [14, 248, 249]. These limitations are attributed mostly to the coverage of the methods used in generating the datasets. For instance, interactions based on mass spectrometry predict few proteins involved in transport and sensing whereas yeast two-hybrid generated data fail to cover some categories for example, proteins involved in translation [246]. Traditionally, experimental datasets are cross validated or quality-checked by benchmarking with a reference set of trusted interactions [247]. However, the advent of *in silico* techniques and scoring schemes has helped to quality-check these datasets prior to functional and structural analysis of the interactome.

The importance of weighing functional associations is to control the biases, uncertainty of data and noise associated with experimental methods. The effectiveness of a scoring function used in predicting the reliability and/or confidence in functional associations is critical in managing the

experimental-associated limitations. This is because, the degree of confidence-level of a functional dataset is in direct relation to the reliability of the data. Inability to carefully weigh the functional associations, results in propagation of annotation errors which in turn lead to a compromise of the integrity and stability of the generated network [245].

In this chapter, we leverage a novel effective information-theoretic based functional scoring scheme presented by Mazandu and Mulder [245] to score the functional associations obtained from sequence BLAST, conserved domains interaction datasets from InterPro as well as other functional interaction datasets. The scheme used in our study have been tested to produce a reliable functional network with higher coverage [245]. In relation to scoring functional pair-wise relationship, this method provides the ability to modify parameters based on the users confidence in the data source. On the other hand, this method considers not only the number of common signatures shared by two proteins in a protein domain and family data, but also considers the nature as well as databases and experiments from which the information was retrieved.

3.3.1 Scoring InterPro Datasets

Similarity score (X_{ij}) between a protein pair (p_i) and (p_j) with common signatures (S_k) is measured by the minimum number of the occurrence of these signatures [245]. This is defined mathematically in **Equation 1** below,

$$X \equiv X_{ij} = \sum_{k=1}^{M} \min\{n_{ki} n_{kj}\} \tag{1}$$

where n_{ki} and n_{kj} represent the number of occurrences of signatures and k is the number of proteins starting from 1 to the last(M).

Due to the associated level of uncertainty in experimental datasets, they naturally follow a normal distribution when compared to other distributions. This implies that, the datasets can be summarized by its mean (μ) and standard deviation (σ). Also, the optimal distribution maximizes information entropy in the dataset. Information entropy is a measure of uncertainty within a dataset. It is defined as the average rate at which information is produced by a stochastic source of data.
Therefore, as established by Mazandu and Mulder [245], the confidence level (δ) of the similarity score (X) implemented in the customary python algorithm for scoring protein family and domain is defined as shown in **Equation 2**

$$\delta \equiv \delta(X, \sigma, \alpha) = \phi\left(\frac{X^{\alpha}}{\sigma}\right) \tag{2}$$

where ϕ is the cumulative probability function of a normal distribution, α is the calibration control parameter which strengthens the impact of the confidence level.
The scoring scheme rectifies the dataset to remove all outliers, thus maintaining data points that lie within a normal distance. After rectifying, the information entropy related to the dataset is computed using the binary entropy function shown in **Equation 3**

$$H_2(\delta) = -\delta \log_2(\delta) - (1-\delta)\log_2(1-\delta) \tag{3}$$

The scheme computes the functional relationship score between protein pairs sharing common signatures. The functional relationship score is defined as shown in **Equation 4**

$$\Gamma(\delta) = 1 - H_2(\delta) \tag{4}$$

The reliability or confidence score of the functional relationship between two proteins is defined as shown in **Equation 5**

$$R = \frac{\Gamma(\delta)}{\max\limits_{s} \Gamma(s)} \tag{5}$$

Upon running the customary python scripts on the InterPro datasets, the pair-wise interaction derived comprised of 1,013 interactions among 256 unique proteins. The functional interaction score ranged from 0.4 to 1.0.

3.3.2 Scoring Protein Sequence Similarity

The scoring scheme presented uses the bit score between pair-wise homologous sequence alignments (s_1, s_2). The bit score, denoted as $(S(s_1, s_2))$, provides a mean for defining homology between pair-wise sequences by measuring the average information or features available per amino acid position in the aligned pair-wise sequence [245]. Homology is simply defined as common evolutionary ancestry between genomic sequences [250].
The bit score is obtained through pair-wise homologous sequence BLAST [244]. BLAST sequence similarity estimates the bit score by identifying common features and estimating statistically significant similarity that reflects shared common ancestor [250]. Also, the scheme presented uses the mutual information $I(s_1, s_2)$ between pair-wise homologous sequence alignments (s_1, s_2). Mutual information is the underlying common substantial biological features contained in pair-wise homologous sequences [250]. This information is based on the fundamental postulate about homologous sequences which is paraphrased as "the closer the similarity between a protein sequence pair, the closer in evolution" [251]. This therefore implies that, the bit score is in direct relation with the mutual information between a protein sequence pair. Therefore, Mazandu et al. [245] established the relationship shown in **Equation 6** below.

$$S(s_1, s_2) = \lambda I(s_1, s_2) \tag{6}$$

where λ is a constant defining the relationship.

The concept of homology is fundamental to *in silico* analysis of both DNA and protein sequences. However, the ability to establish homology when two sequences have more mutual information other than would be expected by chance is critical for the analysis [250]. This is because homologous sequences do not always share significant sequence similarity [250]. For example, some homologous protein alignments are not significant but these proteins are characterized as homologous based on statistical significant strong sequence similarity to intermediate sequence [250]. For this reason, the scheme used measures the mutual biological evolution information [251] available per amino acid position to distinguish an alignment from chance using the relative entropy of target residue and background distributions shown as **Equation 7**

$$H(s_1, s_2) = \sum_{i,j} q_{ij} s_{ij} = \sum_{i,j} q_{ij} \log_2 \left(\frac{q_{ij}}{q_i q_j} \right) \tag{7}$$

where q_{ij} is the target *residue* substitution frequency which is defined as the probability of finding a residue i aligned with residue j after a certain amount of evolution given that they both evolved from common ancestor who had residue k at that position. q_i is defined as the probability of occurrence of a residue i in a set of sequences. s_{ij} is the similarity score between residue i and j.
The reliability score for the pair-wise sequence similarity implemented in the algorithm used is defined as **Equation 8**

$$R(s_1, s_2) = \frac{I(s_1, s_2)}{\max\{H(s_1), H(s_2)\}} \tag{8}$$

where $H(s)$ is the relative entropy obtained after aligning protein sequence s by it self.

We used our generated *Plasmodium falciparum* sequence BLAST data as input data for a custom python script incorporating the computations described above. The generated output functional interaction data comprised of 231 interactions between 130 proteins.

3.3.3 Scoring High-throughput Experimental Datasets and Interologs

In the analysis for this section, the following criteria was set in order to score pair-wise functional associations of experimental and interolog datasets retrieved from databases and literature. Interactions in interolog datasets are based on the concept that orthologs of pair-wise proteins should also have functional interactions [252]. The criteria listed below were fundamentally based on supporting evidence, herein referring to experiments, databases and/or reported literatures confirming such functional interactions.

1. The number of experimental methods that have confirmed such functional interaction.
2. The number of databases that have reported such functional interaction.
3. The number of times the functional interaction have been reported in literature.

Experimental functional interaction datasets generated by Wuchty et al. [106], LaCount et al. [94] and IntAct database [109] as described in **Table 4** were scored based on these criteria. Pair-wise functional interactions within each dataset supported by one evidence was assigned a reliability score of 0.4. On the other hand, we assigned a reliability score of 0.7 if the functional interaction is supported by two or more evidence be it from literature, databases or experimental methods. Overall, we were able to define reliability score for these dataset in order to perform filtering and further integrative analysis. On the contrary, datasets including Wuchty et al. [103, 104, 105] which had reliability scores were maintained for filtering and integrative analysis.

3.4 Overall Filtering of Datasets

Most of the datasets used in this chapter (summarized in **Table 9**), had different gene identifiers or protein identities (IDs) including but not limited to Ensembl and IntAct IDs depending on the source. To ascertain a uniform ID for convenient data manipulation and downstream analysis, we mapped all IDs unto Uniprot database to retrieve their corresponding Uniprot IDs. Genes or proteins which had no corresponding Uniprot ID as at the time of this analysis were discarded. However, upon careful analysis, such proteins were uncharacterized.
Unlike other studies which use both reviewed and unreviewed or uncharacterized interactions, our study sought to extract from each dataset, pair-wise functional interactions between *Plasmodium falciparum* isolate 3D7 manually annotated proteins.
Table 9 describes individual datasets and the number of pair-wise interactions between manually annotated proteins whereas **Figure 6** shows the graphical representation. The obtained annotated interactions were used to generate the unified pathogen network.

Table 9. Extracted functional interactions between manually annotated *Plasmodium falciparum* isolate 3D7 proteins.

Interaction source	Number of reported interactions	Number of reported proteins	Number of observed reviewed proteins	Number of observed pair-wise interactions between reviewed proteins	Reference
LaCount et al.	2,864	1,310	62	17	[94]
Wuchty et al. 1	19,979	2,321	85	74	[104]
Wuchty et al. 2	993	323	12	11	[103]
Wuchty et al. 3	5,459	1,988	91	32	[106]
Wuchty et al. 4	4,918	1,872	81	15	[105]
IntAct	2,916	1,343	67	26	[109]
InterPro	1,013	256	98	241	[98]
Scored BLAST sequence similarity	1,090 (BLAST)	163	130	231	[245]
STRING	617	163	114	386	[102]

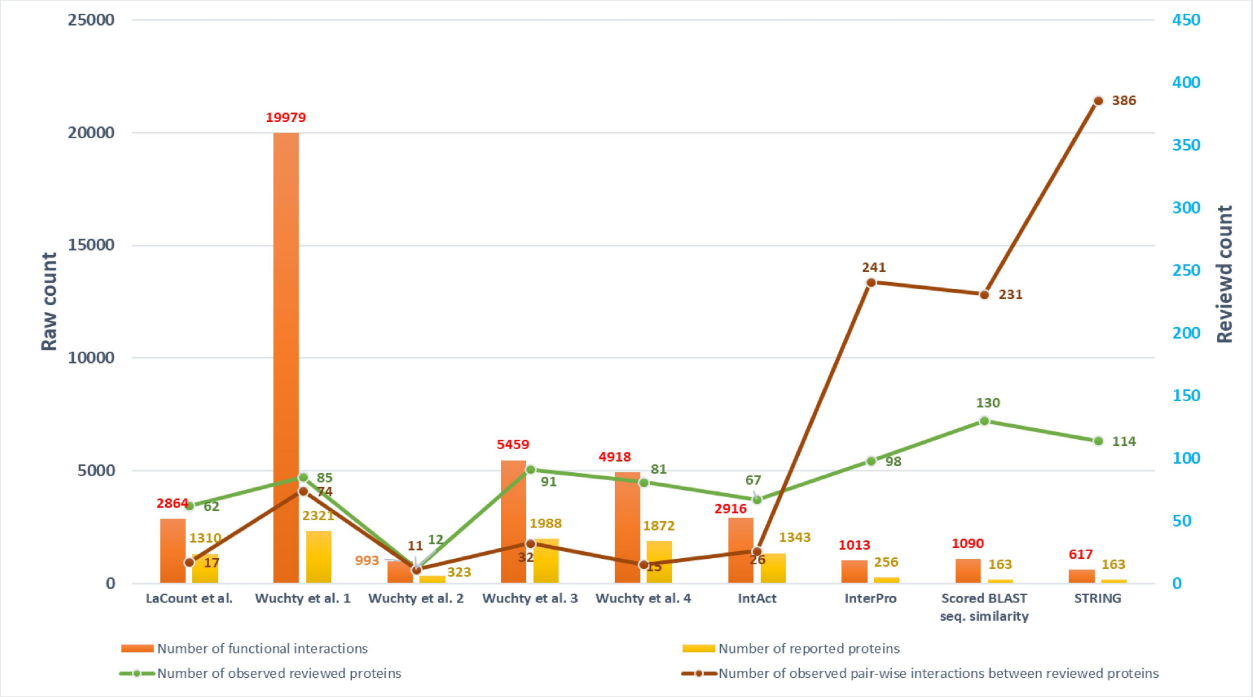

Figure 6. Graphical representation of the extracted functional interactions between reviewed parasite proteins from various datasets as described in **Table 9**.

3.5 Constructing *P. falciparum* Functional Network

At this section we integrated all the extracted parasite datasets (**Table 9**) into a unified functional PPIN. The datasets were integrated using customary python scripts that operated mainly on the networkX package [253]. The generated functional network comprised of 799 interactions between 155 reviewed *Plasmodium falciparum* isolate *3D7* proteins with an average degree of 10.3097. The functional network includes chloroquine resistance transporter (*Q8IBZ9*), ornithine aminotransferase (*Q6LFH8*), cofilin/actin-depolymerizing factor homolog 1 (*Q8I467*), T-complex protein 1 subunit eta (*O77323*), anamorsin homolog (*C0H4X5*), phosphoglycerate kinase (*P27362*), adenylosuccinate synthetase (*Q8IDF8*), 40S ribosomal protein S3a (*O97313*), 60S acidic ribosomal protein P2 (*O00806*), v-type proton ATPase catalytic subunit A (*Q76NM6*), and proliferating cell nuclear antigen (*P61074*) proteins targeted by artemisinin [254]. These proteins are involved in the parasite's *protein biosynthesis, glycolisis, antioxidant defense system, hemoglobin digestion* and *immune response pathways* [254]. However, *pfcrt* has been shown to contribute to artemisinin resistance.
662 functional interactions with confidence score $\geqslant 0.3$ between 140 nodes were selected for further downstream analysis such as structural and functional analysis. The network consist of eight clusters which are shown as different colours in **Figure 7**.

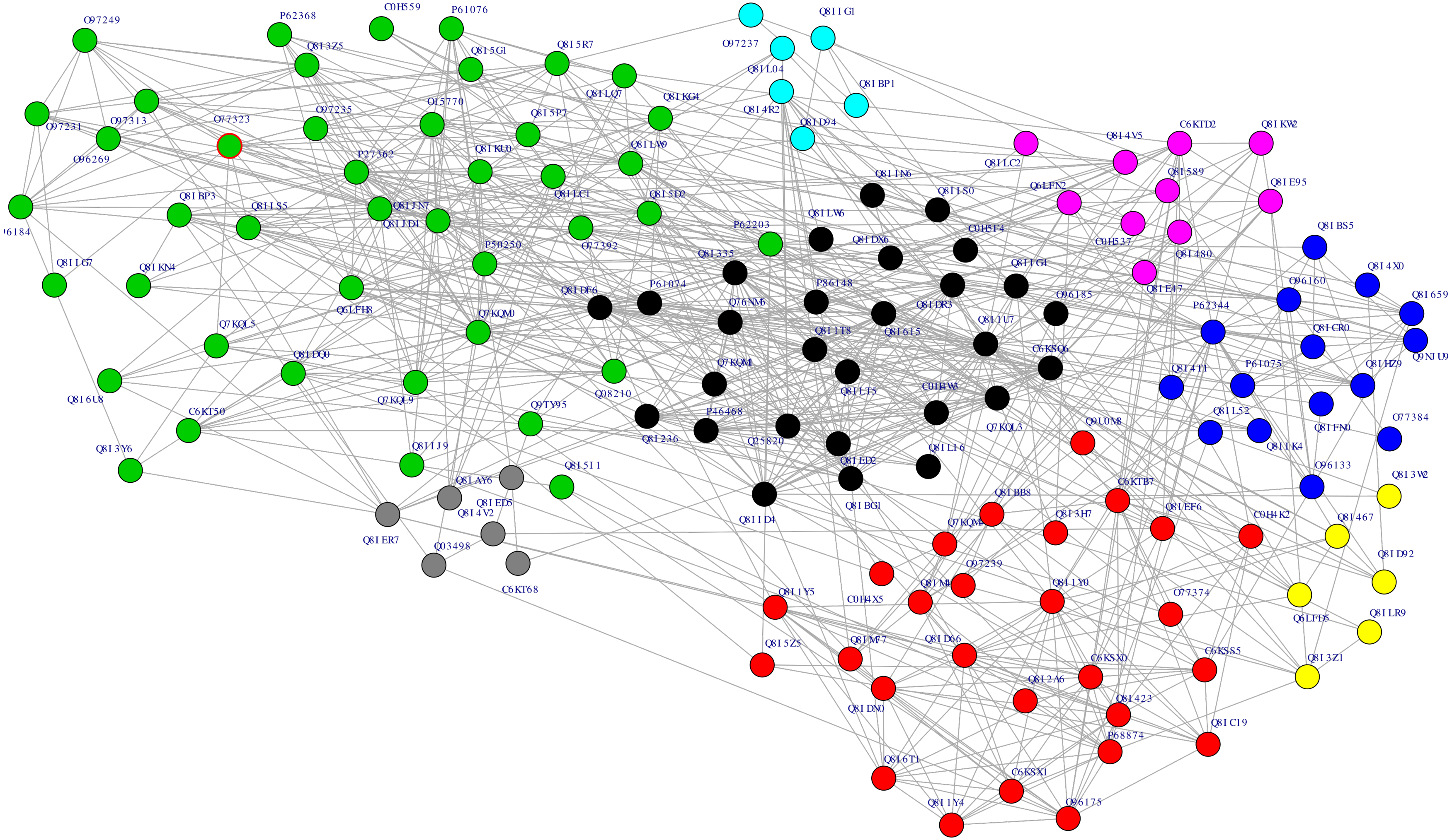

Figure 7. A graph network of *Plasmodium falciparum* protein-protein interactions between reviewed proteins. The nodes are coloured according to clusters or subnetworks whereas the edges are shown as lines. The subnetwork with black nodes is the most key hub in the generated functional network.

3.6 Structural Analysis of *P. falciparum* Proteome Functional Network

PPIs are studied experimentally using biochemical, biophysical and genetic techniques [246]. However, in this post-genomic era, computational approaches to analyzing protein interaction have come to complement high-throughput interaction-detection methods.

PPINs are characterized by topological and dynamic properties essential for biological activities at the molecular and systems level. The term topology herein simply refers to arrangement of nodes and edges within the network. These complex interactomes are distinguished by overlapping dense clusters or subnetworks. The overlapping interconnected nodes suggest the presence of a well defined functional and topological core of the network [14, 104].

PPINs are represented graphically as nodes and edges. The nodes represent proteins whereas edges or links represent the information connecting the nodes. Careful study of protein-protein interactions (PPIs) is very critical for understanding cell physiology in normal and disease state thus, serving as a crucial tool for leveraging disease mechanisms and identifying essential proteins that can serve as drug targets. However, the ability to extract meaningful information from these networks requires a systematic computational analysis of the complex topological features of the network to investigate the impact of each node interaction in relation to the integrity and stability of the system. This would contribute to identifying dynamic patterns of essential information flow within the network for predicting drug targets.
In the following section, we described the application of centrality metrics to analyze the network and predict key nodes (proteins).

3.6.1 Computing Topological Centrality Metrics

Centrality metric provides a quantitative measure of the functional significance of a node within a network. It is a measure of the influence a node(gene,protein,etc) plays within a network or the ability of a particular node to be influenced by other nodes within the same network. The weight of this metric is determined by the node's connection topology.
We computed degree, betweenness and closeness centrality metric using NetworkX package in python. These metrics were measured to evaluate the topology of nodes and edges within the network. This would facilitate elucidating essential nodes characterized by many connections or hubs thus, contributing to identifying essential processes within the system.
Degree metric is a topological property of a network that measures the ability of a node to interact or communicate directly with neighbouring nodes [252]. The degree distribution within a network is a measure of the scale–free property of the functional network. Nodes with higher degree (above average) are central to the connecting nodes within the network to enable connection at short steps. Higher degree nodes form the degree-based subnetwork or hub. These type of subnetworks are specific to particular regions of the network. The degree metric is represented as shown in **Equation 9**

$$degree(p) = \sum_{q \in \mathcal{N}} \delta(p, q) \tag{9}$$

where p and q are proteins

$$\delta(p,q) = \begin{cases} 1 & \text{if protein q is functionally linked to protein p} \\ 0 & \text{otherwise.} \end{cases}$$

Closeness measure determines nodes that are relatively closer to all nodes in the network [97, 248, 252]. It is measured by the ability of a node to access information from other nodes. It is given by

Equation 10

$$C(p) = \frac{|L_c| - 1}{(n_C - 1) * S_r(p)} \tag{10}$$

where
n_c is the number of nodes in the path of a node of interest.
$|L_c|$ is the number of functional interactions connecting the nodes.

Betweenness centrality is a measure of the influence that a node has over the flow of information within pair of nodes in the network [97, 248, 252]. It is based on the idea that flow of information is between shortest paths connecting node pairs. This implies that nodes with high betweenness are very essential and regulates the functioning of the system. In view of that, knocking out such nodes with higher betweenness would significantly interfere with the activities of the pathogen linked to its survival. Nodes with higher betweenness form the structural subnetwork or hubs ensuring the flow of information within the network thus maintaining the integrity of the network. Unlike the degree-based hub, structural hub might have connections within several degree-based hubs [252]. Therefore, removal of a structural hub significantly disintegrates the network.
The betweenness metric is represented as shown in **Equation 11**

$$B(c) = \sum_{(a,b) \in N_c} \frac{\sigma_{ab}(c)}{\sigma_{ab}} \tag{11}$$

where σ_{ab} is the shortest paths between protein a and b passing through protein c.
Shortest path between node pairs in a biological network is that path with the minimum number of edges.

For each predicted pair-wise interaction, its degree, betweenness and closeness is computed to identify highly connected nodes.

3.6.1.1 Assessing High-Degree Proteins of the Obtained Parasite Network

Computing centrality metric as described above using a developed python-based algorithm revealed that the degree of nodes within the parasite's functional network ranged from 1 to 33 with an average score of 9.243. The degree score describes the number of direct functional interaction with each node having at least 9 direct functional interaction as shown by the average degree score. *Adss* gene with uniprot ID Q8IDF6 (Adenylosuccinate synthetase) had the highest degree of 33. This protein is involved in salvage pathway for the synthesis of purine nucleotide [93]. Also, 16 nodes had degree of 1. **Figure 8** shows the node degree distribution which presents an overview of each node and the number of associated direct functional interaction with other nodes within the network. **Figure 9** shows the path length distribution of the nodes within the *Plasmodium falciparum* network. The path length describes the shortest paths between all node pairs within the network, which indicated the level of information spread across the network [252]. **Table 10** provides a summary of the network parameters of the parasite's functional network.

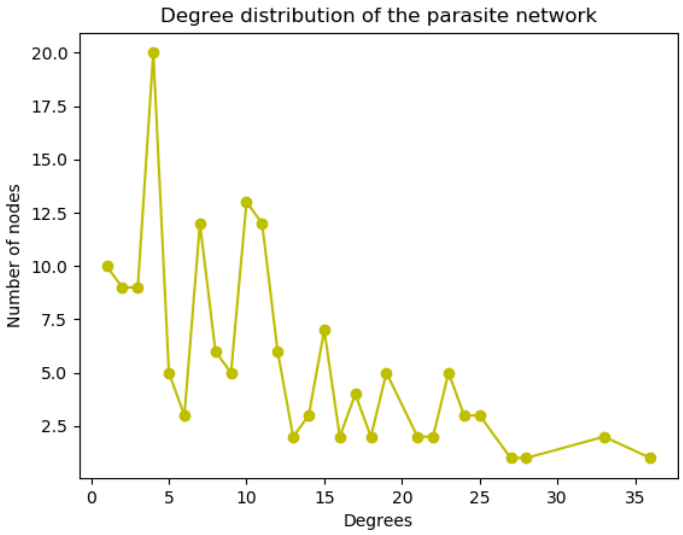

Figure 8. A graph showing the degree distribution of the various nodes in the functional network, indicating the scale–free property of a network whereby few nodes are characterized by high degree.

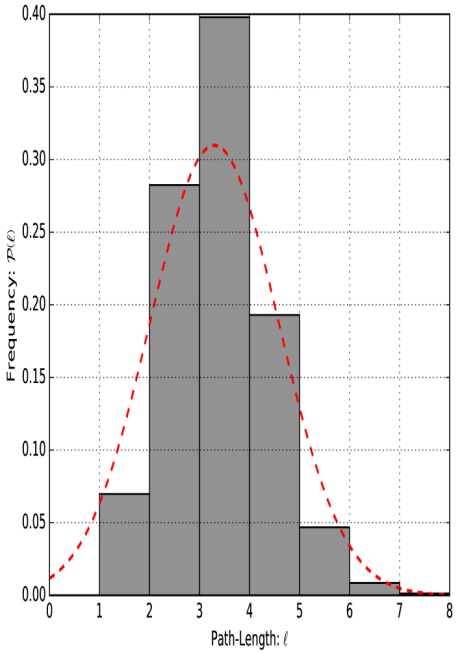

Figure 9. A bar graph showing the path length distribution between pair-wise proteins (nodes) in the parasite's network with the minimum, maximum and average length been 1, 7, 2.89577 respectively. The average length is the mean of all the shortest paths between paired proteins and it is a measure of information relay within the network.

3.6.1.2 Betweenness

The betweenness score ranged from 0 to approximately 923, with an average score of 122.507 indicating the small world property of a network, a measure of non-neighbouring nodes within the network to influence each other through indirect functional interaction. From our results (**Table 10**)

PCNA gene with uniprot ID P61074 (Proliferating cell nuclear antigen) had the highest score. The protein is involved in the control of eukaryotic DNA replication by increasing the polymerase's processibility during elongation of the leading strand [93]. 29 proteins had the lowest betweenness score of 0. **Figure 10** describes the relationship between the degree and betweenness score of nodes in the parasite network. It describes nodes that have the ability to directly and indirectly establish functional interaction within the network.

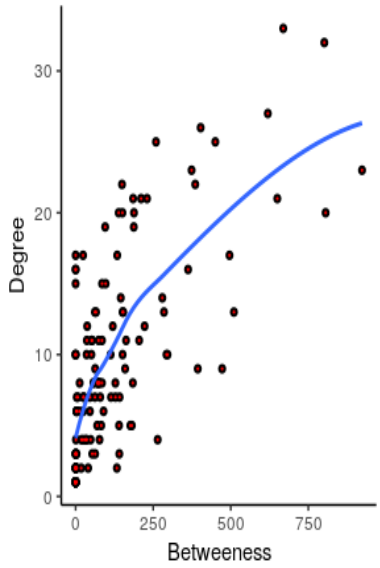

Figure 10. Relationship between the degree and betweenness centrality measure of the nodes (dots) in the parasite network. It is observed that majority of nodes have betweenness score between 0 and 250 with maximum degree of about 20, suggesting the small world property of the network whereby nodes that are not neighbours within the network can interact through other nodes.

3.6.1.3 Closeness

The closeness score ranged from 0 to 0.45808 with an average score of 0.331. *PCNA* gene had the highest closeness score. **Figure 11** and **Figure 12** describes the relationship between closeness–betweenness and closeness–degree metric respectively. It is observed that some nodes have both high closeness and betweenness score. This suggest that for some non–neighbouring nodes that indirectly interact, the more closer the nodes are to each other the higher the probability of influencing and relaying signals or information. However, the converse is true for nodes characterized by both low closeness and betweenness score. In addition to that, some non-neighbouring nodes are characterized by high closeness score but low betweenness score. This may also suggest that, although the pair of non–neighbouring nodes are closer but do not influence each other. This might be that, the nodes contribute together in few processes as compared to high closeness–high betweenness nodes. **Table 10** describes the summary of the parasite's unified network properties.

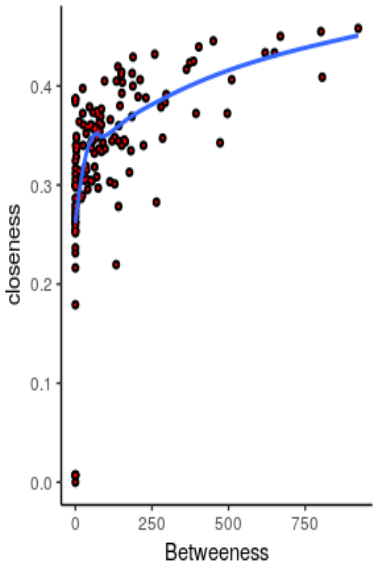

Figure 11. Relationship between the closeness and betweenness centrality measure of the nodes (dots) in the parasite network, suggesting that some nodes are characterized by either high closeness and betweenness score, high closeness low betweenness score or low closeness low betweenness score.

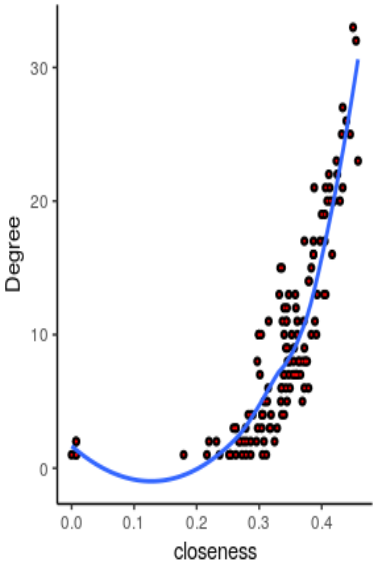

Figure 12. Relationship between the degree and closeness score of nodes in the parasite network.

Table 10. General *Plasmodium falciparum* functional network parameters

Parameters	Value
Number of nodes (proteins)	140
Number of functional interactions	662
Maximum Degree	33
Minimum Degree	1
Average Degree	9.243
Maximum Betweeness	923.08
Minimum Betweeness	0
Average Betweeness	122.507
Maximum closeness	0.458
Minimum closeness	0
Average closeness	0.331

3.6.2 *Plasmodium falciparum* Selective Variant Network

Malaria selective variants are those genes known to be involved in selective events that hinders the progress in malaria control such as reduced parasite clearance rate and malaria–drug resistance [255]. Usually, the contribution of these variants to a selective event is population–specific based on the pathogen's genetics, suggesting that variants common is different populations may not be under selection in each population, but may contribute to a selective event [255]. This implies that, a functional network comprising of selective variants would enable us investigate the functional interactions between these variants thus, helping to investigate the combinatorial effect of these genes within a system towards a specific selective events.

In this section, *Plasmodium falciparum* known selective variants and other reported variants expressing strong signature of selection were retrieved from databases and literature. We queried these variants as input data in STRING database to retrieve a malaria selective variant–specific functional network. The malaria–specific interactome consisting of 206 interactions between 84 genes was generated as the output from STRING. The generated network was filtered to include only genes or proteins specific to *Plasmodium falciparum* isolate *3D7*.

From the generated filtered network, we focused on interactions between some artemisinin targets [254] as shown in **Figure 13**. The purpose of this analysis was to investigate the functional interactions to help determine the level of association between the selective variants and drug targets in order to explore potential patterns that could underly drug resistance.

Among these variants are *dhps* (Q8IAU3), *pfmdr1* (PFE1150w), *plasmepsin2* (PF14_0077), *dhfr* (Q8I1R6) and *pfcrt* (Q8IBZ9) known genes with associated mutations executing specific biological functions conferring resistance to antimalarial drugs. These variants served as the main nodes connecting the clusters.

Our results revealed a higher level of functional interactions between selective variants associated with drug resistance. Analysis on the shortest paths between the drug targets and the selective variants to explore routes for possible resistance development suggested that, resistance to artemisinin drug targets is likely to involve contributions from these resistant genes and might not follow the mode of

resistance to chloroquine and sulfadoxine-pyrimethamine characterized by *pfcrt*, *dhps* and *dhfr*. Due to the difference is functional interactions (direct or indirect) between the targets and the selective variants, there is a higher likelihood that, the targets will experience varying degrees of resistance.

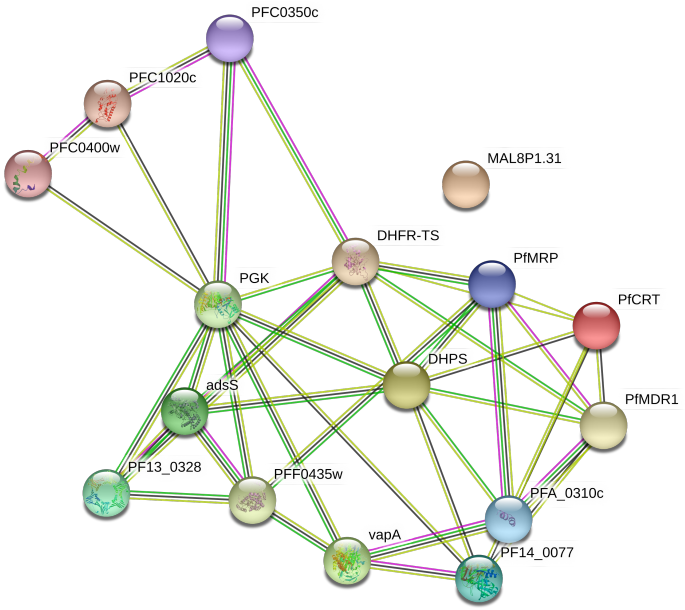

Figure 13. Functional interactions between malaria-resistance conferring genes and some artemisinin drug targets.

3.6.3 Network Protein Clustering

Network clustering involves the decomposition of networks into sub-networks or communities of highly interconnecting nodes. PPI networks are undirected networks that are characterized by their modularity, and as such, they have the likelihood of node clustering [252]. The partitioning of a network is measured by the transitivity or clustering coefficient. The transitivity of a network is a measure of the degree to which the relationship between two connecting nodes within a network is transitive.

In that regards, the purpose of clustering the generated unified network is to identify hubs or densely connected nodes forming subnetworks to facilitate extraction of critical functional nodes. This process helps to identify essential biological mechanisms underlying the system. Various algorithms including but not limited to network division algorithms which have the capacity of detecting inter-community links within a network [256, 257, 258] and agglomerative algorithms which merges similar nodes within a community [259] have been developed for network clustering. Also, optimization methods have been developed for such network clustering [257, 260].

However, we implemented a simple but powerful algorithm described by Blonde et al. [113] for network clustering. This is because, it has been investigated to show that it out-performs other methods such as divisive algorithms especially in terms of computational time and output quality [113]. The algorithm herein finds high modularity partitions of large networks within shorter time and presents a complete hierarchical hubs or subnetworks [113]. The ability to elucidate clusters, key proteins and characterize associated genes was enhanced by the centrality scores of the nodes as well as mapping disease-associated genes unto the unified parasite network.

Nodes with average betweenness and closeness score with their corresponding degree score were considered as key targets because of their significant influence within the network. Further structural and functional analysis on the 43 key genes or nodes helped to filter out key candidate targets or proteins highly essential to disease pathogenesis and network integrity. These key candidate targets

are more central and influential in the network.

The generated network (**Figure 7**) consisted of 8 clusters of which 5 contained the 43 key proteins. 2 of the 5 subnetworks contained candidate key proteins as described in **Table 11**. The key candidate proteins were identified in cluster 2 and 4. Cluster 2 contained 1 key candidate gene encoding key protein, Putative E3 ubiquitin-protein ligase protein *PFF1365c* (C6KTB7). There is accumulation of evidence that C6KTB7 is potential candidate for malaria vaccine and drug development [93, 261, 262]. C6KTB7 is involved in the protein ubiquitination pathway of the pathogen [93]. Studies have shown that many biological processes and substrates are targeted by the ubiquitin pathway such that instability or modification in ubiquitination and deubiquitination reactions influences the pathogenesis of many eukaryotic system related diseases [261]. For instance, dysregulation of ubiquitin ligase is associated to neurodegenerative disorders such as Parkinson's disease and infectious diseases including tuberculosis [262]. This is usually associated with interference with immune response.

This protein significantly influences the parasite's development and malaria pathogenesis [263]. This is because, it regulates various cellular process and pathways critical for the pathogen's survival in the human host. For example, it is responsible for positive regulation of DNA-templated transcription and epigenetic factors such as histone H3-K4 methylation, essential for transcription regulation [261]. Interestingly, several studies have shown that, inhibition of the activities of C6KTB7 and ubiquitin proteasome system is essential for many disease treatment including *Plasmodium falciparum* malaria [261, 263]. The functional interactions of C6KTB7 in the unified parasite network is shown in **Figure 14**.

Cluster 4 contained 2 key proteins of which one was a candidate key protein expressed in the parasite's erythrocytic developmental stage. The 2 candidate genes are *PF07_0086* (Q8IBP1) and *PFF1440w or SET1* (C6KTD2). These genes have also been reported to be potential candidate genes for an effective malaria vaccine [264]. However, in our study, putative histone-lysine N-methyltransferase 1 (C6KTD2) emerged as one of the 2 candidate key proteins critical for disease pathogenesis after further structural and functional analysis. This gene is expressed in the merozoite stage of the parasite development. C6KTD2 is known to play an essential role in chromatin structure and gene expression in the parasite [265]. Also, it is mainly involved in histone lysine methylation process which usually involves the synergistic effect of histone-lysine methyltransferases and histone lysine demethylases [265].

The cluster formed by C6KTD2 in the parasite network is shown in **Figure 15**. **Table 12** describes the degree, betweenness and closeness centrality score for C6KTB7 and C6KTD2 within the pathogen's functional network.

Table 11. Classification of the unified parasite network into subnetworks or hubs.

Cluster ID	Number of proteins	Number of key proteins	Number of candidate proteins	Number of candidate key proteins
0	24	6	0	0
1	27	17	0	0
2	23	3	1	1
3	40	16	1	0
4	16	16	2	1

Table 12. Degree, betweenness and closeness centrality score of C6KTD2 and C6KTB7 within the parasite unified functional network.

Uniprot ID	Gene name	Description	Betweenness	Degree	Closeness
C6KTD2	*SET1*	Putative histone-lysine N-methyltransferase 1	510.18	13	0.40623
C6KTB7	*PFF1365c*	Putative E3 ubiquitin-protein ligase protein PFF1365c	284.70	13	0.34726

Also, we observed intersecting nodes within the subnetwork formed by C6KTB7 and C6KTD2. The nodes P62344 and Q81E47 (yellow nodes in **Figure 16**) connected the clusters formed by the parasite candidate key proteins in the parasite network).

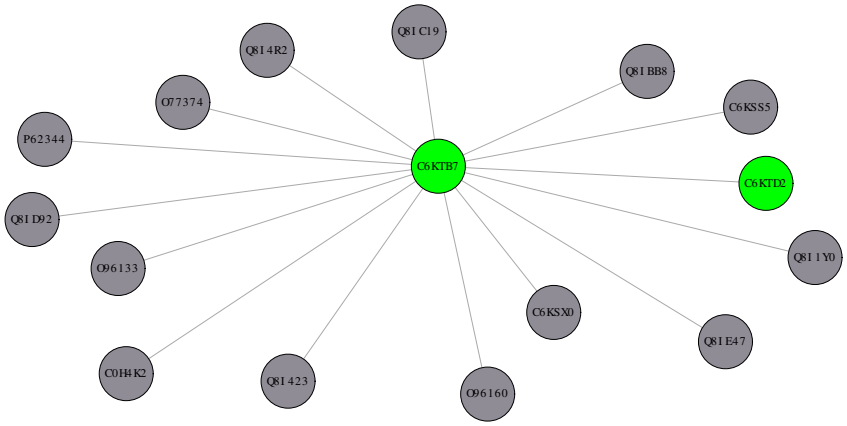

Figure 14. Functional interactions between C6KTB7 (central green node) and directly connected proteins (grey nodes) in the pathogen functional interaction network.

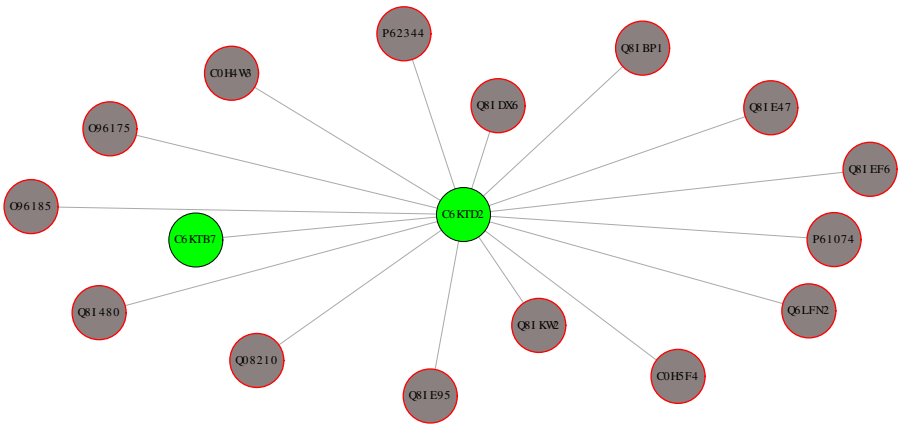

Figure 15. Functional interactions between C6KTD2 (central green node) and directly linked proteins (grey nodes) in the pathogen functional network.

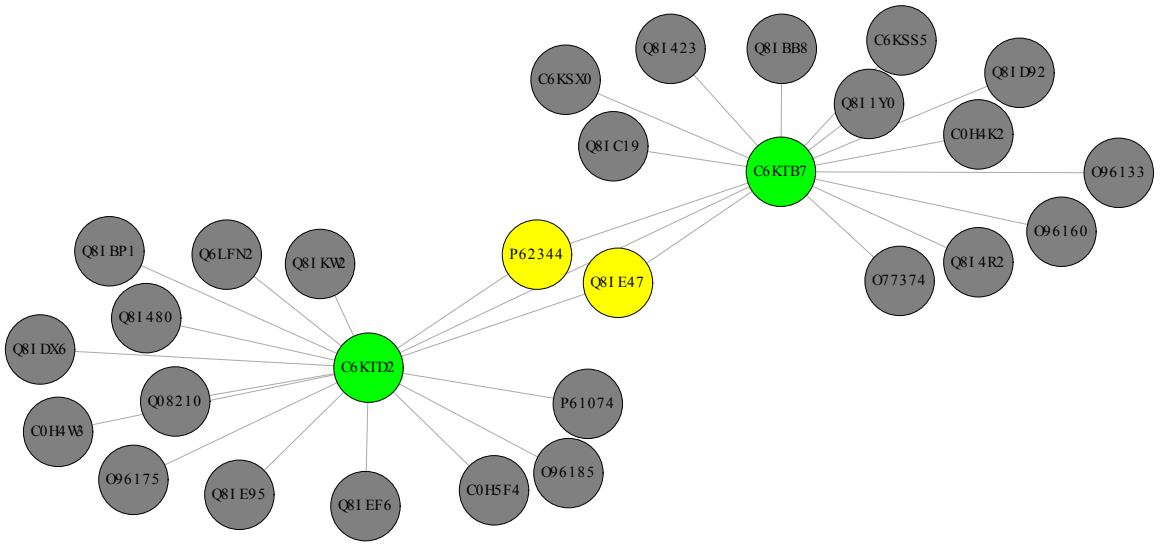

Figure 16. Functional interactions between C6KTD2 and C6KTB7 clusters in the unified pathogen functional network predicted to be involved in development and disease pathogenesis.

3.7 Functional Analysis of Parasite Disease–Associated–Candidate Genes Encoding Key Proteins

In this section, we performed gene annotation and enrichment analysis on disease associated genes (C6KTB7 and C6KTD2) identified during network clustering. We investigated statistically significant gene ontology process or enriched processes which the genes are involved. C6KTB7 is mainly involved in ubiquitin–protein transferase activity through the *protein ubiquitination and modification pathway* (UPA00143, GO:0016567) [93]. The ubiquitylation process, consisting of a complex network of enzymes (E1 ubiquitin-activating enzyme, an E2 ubiquitin-conjugating enzyme and an E3 ubiquitin ligase) within the parasite is known to be highly critical for numerous cellular processes contributing to the parasite's survival and propagation [262]. This usually happens as a result of post-translational modifications within the biological system through processes such as transcriptional regulation and cell cycle progression [262].

C6KTD2 is involved in *histone H3-K4 methylation post–translational modification process* (GO:0051568), transcriptional regulation, activation and silencing [93, 266]. The methylation process regulates various biological processes particularly, it is highly influential in the parasite's ability to invade the red blood cells as well as regulates virulence genes in the parasite [266].

3.8 Summary

In this chapter, we constructed a unified parasite functional network from *Plasmodium falciparum 3D7* isolate reviewed proteins and various functional interaction datasets retrieved from literature and databases. We computed the network centrality scores for the nodes to investigate subnetworks and key nodes critical for *Plasmodium falciparum* malaria pathogenesis. We investigated on nodes with degree above the average degree score of approximately 9 (degree-based subnetwork or hub) and nodes critical in disconnecting functional interactions within the network (structural subnetwork or hub) to identify key candidate targets at parasite side. Importantly, we identified C6KTD2 and C6KTB7 as key candidate target proteins within the parasite functional network for new drug development from structural and functional analysis of the network. Our results consolidates several studies which have predicted these targets for novel drug development using our algorithms, models and network analysis techniques. The functional interaction between these targets may suggest them

as co-targets in combinatorial drug design. We performed functional analysis on the candidate key genes to elucidate enriched processes and pathways in which they are involved. Interestingly, these targets share a common activity of post-translational modifications in the parasite's life cycle. Also, we investigated the relationship between parasite selective variants and their functional interactions with artemisinin drug targets within the unified parasite functional network. We observed that possible down regulation to these drug targets which could lead to resistance may be influenced by the combinatorial effect of these selective variants as compared to the case of chloroquine and sulfadoxine pyrimethamine resistance.

CHAPTER 4

4 Host Malaria–Specific Proteome Functional Network

The fundamental criteria for selecting disease–associated proteins or genes as drug targets requires the potential target(s) to be essential and indispensable to disease aetiology. For instance, in genetic diseases gene therapy involves identifying genetic variants associated with the disorder. However, with infectious diseases like malaria, the criteria requires understanding the complex interplay between the human host and the pathogen to identify essential proteins or pathways [128, 267].
Human functional network serves as a rich and critical resource for understanding protein organization and functions within a cell.
In this chapter, a computational integration technique is applied to build a comprehensive human protein network based on previously identified malaria risk/resistance genes from GWAS in order to finally construct a unified human-*Plasmodium falciparum* protein network. Primary datasets, such as protein sequences, functional data sets from high–throughput experiments, protein signatures from databases such as InterPro and *in silico* generated functional datasets were used to construct a unified host malaria–specific protein–protein interaction network.

4.1 Functional Interaction Datasets for Constructing Human Functional Network

Similar to datasets implemented in generating a unified interaction dataset for the pathogen (**Chapter 3 section 3.2**), this section uses functional interaction and genomic datasets summarized in **Table 13**. Human protein interaction network proposed by Bossi and Lehner described in **Table 4** was used as one of the datasets to build the functional network. The dataset was generated by integrating tissue-specific interactome and gene expression data from 21 different sources. The network comprises of 80,922 physical interactions among 10,229 human proteins.
In addition, human PPIs derived from reactome database [96] was implemented. Reactome is an open-source, open access, manually curated and peer-reviewed pathway database. The interactome consisted of 79,620 interactions between 8,059 proteins.
20,395 reviewed human protein sequences were retrieved from Uniprot database as at the time of this study to generate pairwise sequence similarity associations using Basic Local Alignment Search Tool (BLAST) specifically blastp, the protein-protein BLAST algorithm.
Also, human protein interaction data including interologs were retrieved from STRING database [102] for this study. The dataset comprised of 11,759,454 interactions between 19,354 proteins. Finally, interaction data was obtained from InterPro database [171]. The dataset obtained comprised of all UniProtKB proteins and the InterPro entries and individual matching signatures. Overall, we were able to assemble a map of 18,830,696 host protein–protein interaction dataset from six independent sources (**Table 13**).

4.2 Scoring Interaction Datasets

Similar to scoring *Plasmodium falciparum* interaction datasets (**Chapter 3 section 3.3**), the same scoring scheme and criteria described above was implemented in scoring human interaction datasets (**Table 13**). The scored interpro interaction data generated from the scheme comprised of 2,646,550 interactions between 17,797 proteins. Also, the scored protein sequence similarity comprised of 3,807,888 functional interactions between 20,395 proteins.

4.3 Filtering Datasets

All protein or gene IDs were mapped unto uniprot to retrieve their corresponding UniProtKB IDs. Similar to the filtering process implemented on the pathogen's interaction datasets (**Chapter 3 section 3.4**), at this section pair-wise interactions between reviewed human proteins was extracted from each dataset. **Table 13** describes individual datasets and the observed pairwise interactions between human reviewed proteins.

Table 13. Extracted functional interactions between manually annotated human proteins.

Interaction data source	Number of reported interactions	Number of reported proteins	Number of observed unique pair-wise interactions between reviewed proteins	Number of observed reviewed proteins	Reference
Reactome	79,619	8,059	19,736	5,029	[96]
Score BLAST sequence similarity	3,807,888 (BLAST)	20,395	143,533	9,611	[244]
InterPro	2,646,550	35,928	231,799	17,797	[98]
Bossi and Lerner	80,922	10,229	54,238	8,416	[95]
STRING	11,759,454	19,354	5,244,655	18,836	[102]
IntAct	456,263	35,770	169,627	16,061	[109]

4.4 Construction of a Human Functional Network

In generating the host malaria–specific functional network, only pair-wise interactions with high confidence score ie. $score \geqslant 0.3$ were considered. NetworkX package in python was the primary package implemented to construct the network. The final human unified network comprised of 4,133,136 functional interactions between 20,329 unique nodes.

4.5 Structural Analysis of Human Proteome Functional Network

Exploring the constructed human protein-protein interactome particularly at the system's level could contribute significantly to elucidate critical biological processes and pathways that influence malaria susceptibility within human. Critical topological analysis helps to unravel nodes that influence the compactness and the information relay capacity within the network. At this section, we performed structural analysis on the network by investigating the topology through degree, betweenness and closeness centrality metric analysis.

4.5.1 Computing Topological Centrality Metrics

The same computations for calculating the centrality metric of the pathogen's functional network was repeated at this section to evaluate the human functional network. **Table 14** describes the summary of the network centrality parameters whereas **Figure 17** shows the path length between the nodes, a

measure of information relay within the network. Comparing **Figure 17** to **Figure 9**, it is observed that the average path length in the host network is shorter due to the difference in network density.

Table 14. Summary of human functional network parameters.

Parameters	**Value**
Number of nodes	20,329
Number of functional interactions	4,133,136
Maximum Degree	3,424
Minimum Degree	1
Average Degree	406.467
Maximum Betweeness	894,567.240
Minimum Betweeness	0
Average Betweeness	13,318.118
Maximum closeness	0.536
Minimum closeness	5e-05
Average closeness	0.436

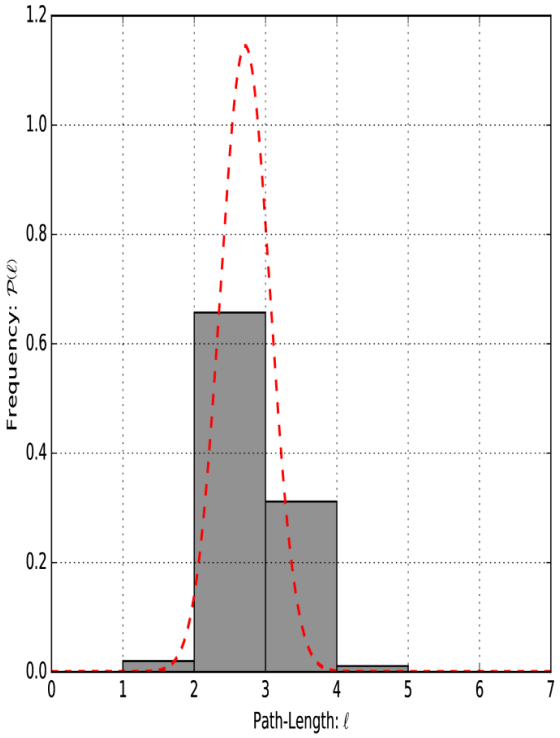

Figure 17. A bar graph showing the path length distribution between the nodes in the parasite's network with the minimum, maximum and mean length been 1, 6, 2.31420 respectively.

4.5.1.1 Assessing High-Degree Proteins

From the centrality metrics computed, it was observed that the degree of nodes within the human functional network ranged from 1 to 3424 with an average degree of 406.467. *LRRK2* gene with uniprot ID Q5S007 (Leucine-rich repeat serine/threonine-protein kinase 2) had the highest degree. The protein is primarily involved in the regulation of autophagy [93]. However, 232 nodes had a degree of 1. Proteins with higher degree contribute significantly to the stability of the network in such a way that they form communities or hubs or subnetworks within the network. Removal of high degree nodes results in disintegration of the connectivity with the network. **Figure 18** describes the degree of distribution of the nodes with respect to the connection frequencies.

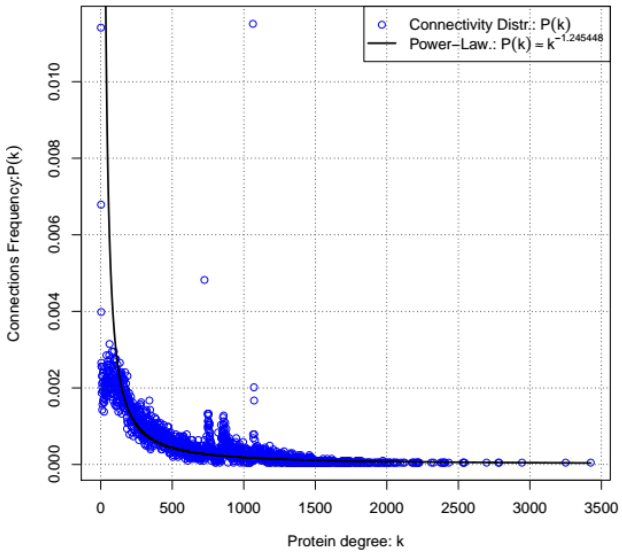

Figure 18. Power law property ($P(k) = k^{-\lambda}$, where λ is the degree exponent) of human functional network generated from integrated heterogeneous datasets. This distribution shows the scale-free property of the functional network whereby large number of nodes are characterized by lower degree score compared to few nodes connected with many neighbours.

4.5.1.2 Closeness and Confidence Measure of a Protein

The closeness score ranged from 5e-05 to 0.53623 with an average score of 0.43554. It was observed that the node with the highest degree had the highest closeness score of 0.53623. However, 8 proteins or nodes had the lowest closeness score. **Figure 19** shows the relationship between the degree and betweenness score of nodes in the human network. It is observed that the closeness of nodes within a network as direct and indirect relationship with the degree such that some nodes characterized by high closeness value could either have a higher or lower degree. Nonetheless, most nodes characterized by high closeness value (≥ 0.2) have degree between 0 and 1000.

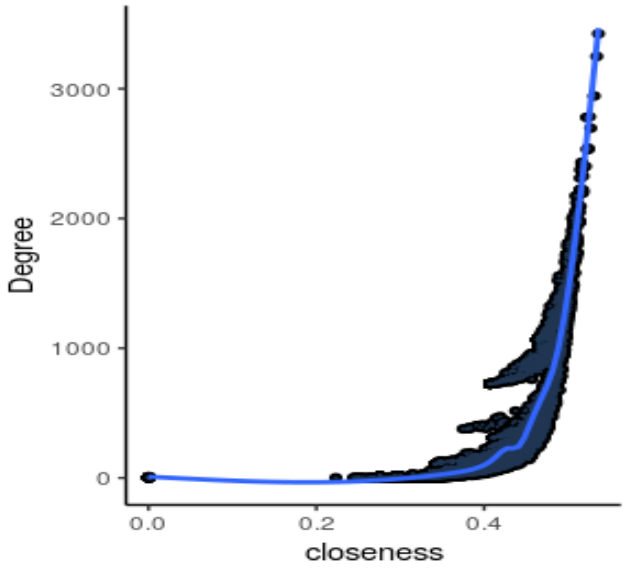

Figure 19. Relationship between the degree and betweenness score of nodes in the human network.

4.5.1.3 Betweenness

The betweenness score ranged from 0 to 894567.24 with an average score of 13318.12. *GAPDH* gene with uniprot ID P04406 (Glyceraldehyde-3-phosphate dehydrogenase) had the highest score of 894567.24. There were 660 proteins with a zero betweenness score. **Figure 20** shows the relationship between the degree and betweenness whereas **Figure 21** describes the relationship between the closeness and the betweenness centrality measure of the nodes in the human network. It is observed that, nodes with higher betweenness score have higher closeness and degree score. This implies that few nodes are associated to relaying wider information thus contributing significantly to the structural integrity of the network.

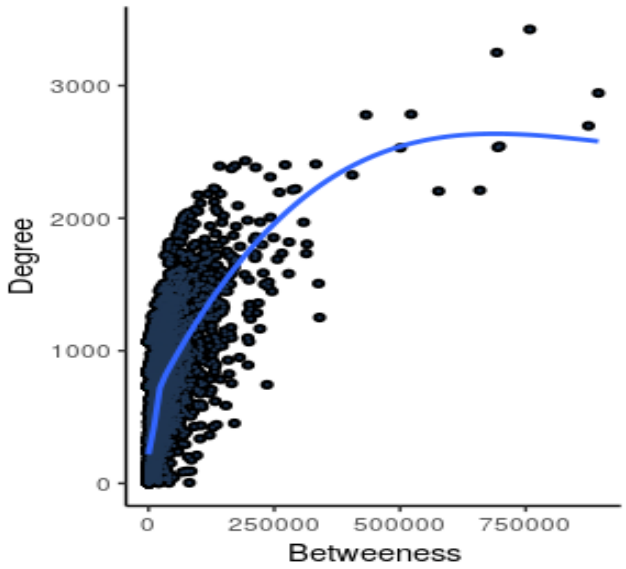

Figure 20. Relationship between the degree and betweenness centrality measure of the nodes (dots) in the human network. It is observed that few nodes with higher betweenness score are characterized by higher degree score.

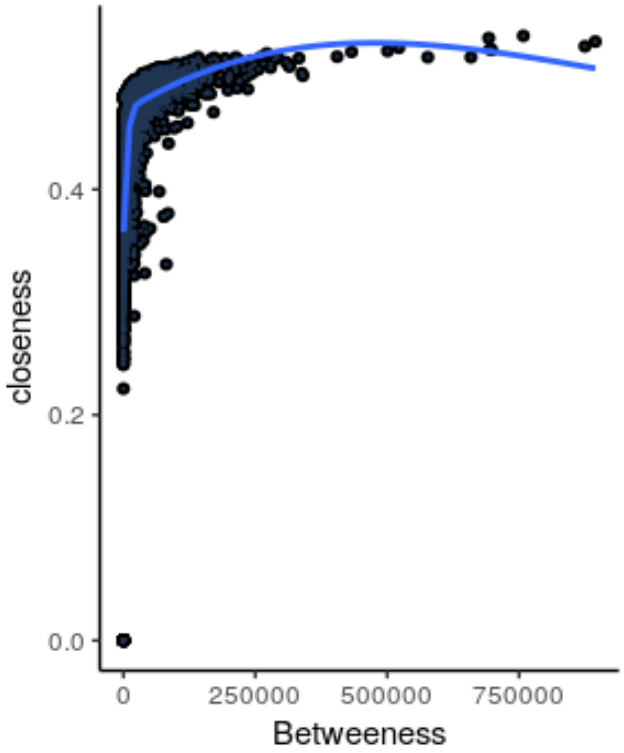

Figure 21. Relationship between the closeness and betweenness centrality measure of the nodes (dots) in the human network.

4.6 Retrieving Disease–Associated Genes from GWAS Summary Statistics Data

Genome–wide association study (GWAS) is a critical single-marker testing technique for identifying and examining disease-specific genetic variants in different populations [92]. In this study, human Malaria susceptibility-associated SNPs for Kenya, Gambia and Malawi populations were retrieved from GWAS summary statistics datasets. **Table 3** describes the number of cases and controls involved in each study.

The homogeneous summary statistics dataset comprised of 20,273,529 SNPs from chromosome one (1) to twenty two (22). A total of 688,577 significant SNPs with p-value $\leqslant 0.05$ were retrieved from the datasets. The SNPs were mapped to genes using the dbSNP database [268]. The dbSNP is a general public archive of all short sequences variant. It contains human single nucleotide variants, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, genomic and RefSeq mapping information for both common variants and clinical mutations. A total of 79 malaria-associated genes (**Table 15**) were retrieved and mapped to uniprot database to retrieve their corresponding IDs. We queried the 79 genes into genemania database [269] to generate a host malaria–specific network shown in **Figure 22**. Functional interaction prediction by genemania is based on co–expression, co–localization, physical interaction, shared protein domains, genetic interactions and literature prediction. **Table A1** in the appendix section describes the genes involved in the functional interaction predicted from genemania database.

Table 15. Malaria-associated genes retrieved by mapping significant SNPs to gene level. The table entails the gene's functional network centrality scores, including betweenness, degree and closeness.

UniProt-ID	Gene name	Description	Betweenness	Degree	Closeness
P23634	*ATP2B4*	Plasma membrane calcium-transporting ATPase 4 (PMCA4) (EC 7.2.2.10) (Matrix-remodeling-associated protein 1) (Plasma membrane calcium ATPase isoform 4) (Plasma membrane calcium pump isoform 4)	13,348.53	295	0.46096
P02818	*BGLAP*	Osteocalcin (Bone Gla protein) (BGP) (Gamma-carboxyglutamic acid-containing protein)	17,285.08	344	0.46416
P16671	*CD36*	Platelet glycoprotein 4 (Fatty acid translocase) (FAT) (Glycoprotein IIIb) (GPIIIB) (Leukocyte differentiation antigen CD36) (PAS IV) (PAS-4) (Platelet collagen receptor) (Platelet glycoprotein IV) (GPIV) (Thrombospondin receptor) (CD antigen CD36)	12,277.06	324	0.46485
P17927	*CR1*	Complement receptor type 1 (C3b/C4b receptor) (CD antigen CD35)	750	61	0.36270
Q16570	*DARC*	Atypical chemokine receptor 1 (Duffy antigen/chemokine receptor) (Fy glycoprotein) (GpFy) (Glycoprotein D) (Plasmodium vivax receptor) (CD antigen CD234)	2,604.91	143	0.42642
P20711	*DDC*	Aromatic-L-amino-acid decarboxylase (AADC) (EC 4.1.1.28) (DOPA decarboxylase) (DDC)	8,905.58	219	0.44328

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P12318	*FCGR2A*	Low affinity immunoglobulin gamma Fc region receptor II-a (IgG Fc receptor II-a) (CDw32) (Fc-gamma RII-a) (Fc-gamma-RIIa) (FcRII-a) (CD antigen CD32)	24,611.42	1,347	0.48942
P08637	*FCGR3A*	Low affinity immunoglobulin gamma Fc region receptor III-A (CD16a antigen) (Fc-gamma RIII-alpha) (Fc-gamma RIII) (Fc-gamma RIIIa) (FcRIII) (FcRIIIa) (FcR-10) (IgG Fc receptor III-2) (CD antigen CD16a)	833.02	1,085	0.46941
O75015	*FCGR3B*	Low affinity immunoglobulin gamma Fc region receptor III-B (Fc-gamma RIII-beta) (Fc-gamma RIII) (Fc-gamma RIIIb) (FcRIII) (FcRIIIb) (FcR-10) (IgG Fc receptor III-1) (CD antigen CD16b)	382.27	1,066	0.46501
P0C091	*FREM3*	FRAS1-related extracellular matrix protein 3	2,648.44	83	0.41347
P11413	*G6PD*	Glucose-6-phosphate 1-dehydrogenase (G6PD) (EC 1.1.1.49)	29,492.51	586	0.47380
P02724	*GYPA*	Glycophorin-A (MN sialoglycoprotein) (PAS-2) (Sialoglycoprotein alpha) (CD antigen CD235a)	198.02	24	0.36143
P06028	*GYPB*	Glycophorin-B (PAS-3) (SS-active sialoglycoprotein) (Sialoglycoprotein delta) (CD antigen CD235b)	832.76	58	0.39935
P68871	*HBB*	Hemoglobin subunit beta (Beta-globin) (Hemoglobin beta chain) [Cleaved into: LVV-hemorphin-7; Spinorphin]	14,739.23	250	0.45899

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P02100	*HBE1*	Hemoglobin subunit epsilon (Epsilon-globin) (Hemoglobin epsilon chain)	10,092.52	178	0.44205
P01889	*HLA-B*	HLA class I histocompatibility antigen, B-7 alpha chain (MHC class I antigen B*7)	4,058.91	1,079	0.46633
P03989	*HLA-B*	HLA class I histocompatibility antigen, B-27 alpha chain (MHC class I antigen B*27)	851.17	1,073	0.46468
P10319	*HLA-B*	HLA class I histocompatibility antigen, B-58 alpha chain (Bw-58) (MHC class I antigen B*58)	451.76	1,072	0.46419
P18463	*HLA-B*	HLA class I histocompatibility antigen, B-37 alpha chain (MHC class I antigen B*37)	332.13	1,071	0.46416
P18464	*HLA-B*	HLA class I histocompatibility antigen, B-51 alpha chain (MHC class I antigen B*51)	332.13	1,071	0.46416
P18465	*HLA-B*	HLA class I histocompatibility antigen, B-57 alpha chain (Bw-57) (MHC class I antigen B*57)	43,292.30	1,075	0.46424
P30460	*HLA-B*	HLA class I histocompatibility antigen, B-8 alpha chain (MHC class I antigen B*8)	27,355.05	1,073	0.46418
P30461	*HLA-B*	HLA class I histocompatibility antigen, B-13 alpha chain (MHC class I antigen B*13)	332.13	1,071	0.46416
P30462	*HLA-B*	HLA class I histocompatibility antigen, B-14 alpha chain (MHC class I antigen B*14)	332.13	1,071	0.46416
P30464	*HLA-B*	HLA class I histocompatibility antigen, B-15 alpha chain (MHC class I antigen B*15)	451.76	1,072	0.46419
P30466	*HLA-B*	HLA class I histocompatibility antigen, B-18 alpha chain (MHC class I antigen B*18)	332.13	1,071	0.46416

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P30475	*HLA-B*	HLA class I histocompatibility antigen, B-39 alpha chain (MHC class I antigen B*39)	332.13	1,071	0.46416
P30479	*HLA-B*	HLA class I histocompatibility antigen, B-41 alpha chain (Bw-41) (MHC class I antigen B*41)	332.13	1,071	0.46416
P30480	*HLA-B*	HLA class I histocompatibility antigen, B-42 alpha chain (MHC class I antigen B*42)	76,201.46	1,330	0.49009
P30481	*HLA-B*	HLA class I histocompatibility antigen, B-44 alpha chain (Bw-44) (MHC class I antigen B*44)	352.06	1,072	0.46425
P30483	*HLA-B*	HLA class I histocompatibility antigen, B-45 alpha chain (Bw-45) (MHC class I antigen B*45)	352.06	1,072	0.46425
P30484	*HLA-B*	HLA class I histocompatibility antigen, B-46 alpha chain (Bw-46) (MHC class I antigen B*46)	451.76	1,072	0.46419
P30485	*HLA-B*	HLA class I histocompatibility antigen, B-47 alpha chain (Bw-47) (MHC class I antigen B*47)	332.13	1,071	0.46416
P30486	*HLA-B*	HLA class I histocompatibility antigen, B-48 alpha chain (Bw-48) (MHC class I antigen B*48)	332.13	1,071	0.46416
P30487	*HLA-B*	HLA class I histocompatibility antigen, B-49 alpha chain (HLA class I histocompatibility antigen, B-21 alpha chain) (MHC class I antigen B*49)	332.13	1,071	0.46416

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P30488	*HLA-B*	HLA class I histocompatibility antigen, B-50 alpha chain (Bw-50) (HLA class I histocompatibility antigen, B-21 alpha chain) (MHC class I antigen B*50)	332.13	1,071	0.46416
P30490	*HLA-B*	HLA class I histocompatibility antigen, B-52 alpha chain (Bw-52) (HLA class I histocompatibility antigen, B-5 alpha chain) (MHC class I antigen B*52)	332.13	1,071	0.46416
P30491	*HLA-B*	HLA class I histocompatibility antigen, B-53 alpha chain (Bw-53) (MHC class I antigen B*53)	451.76	1,072	0.46419
P30492	*HLA-B*	HLA class I histocompatibility antigen, B-54 alpha chain (Bw-22) (Bw-54) (MHC class I antigen B*54)	451.76	1,072	0.46419
P30493	*HLA-B*	HLA class I histocompatibility antigen, B-55 alpha chain (Bw-55) (HLA class I histocompatibility antigen, B-12 alpha chain) (MHC class I antigen B*55)	451.76	1,072	0.46419
P30495	*HLA-B*	HLA class I histocompatibility antigen, B-56 alpha chain (Bw-22) (Bw-56) (MHC class I antigen B*56)	451.76	1,072	0.46419
P30498	*HLA-B*	HLA class I histocompatibility antigen, B-78 alpha chain (MHC class I antigen B*78)	332.13	1,071	0.46416
P30685	*HLA-B*	HLA class I histocompatibility antigen, B-35 alpha chain (MHC class I antigen B*35)	451.76	1,072	0.46419
Q04826	*HLA-B*	HLA class I histocompatibility antigen, B-40 alpha chain (Bw-60) (MHC class I antigen B*40)	332.13	1,071	0.46416

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
Q29718	*HLA-B*	HLA class I histocompatibility antigen, B-82 alpha chain (MHC class I antigen B*82)	332.13	1,071	0.46416
Q29836	*HLA-B*	HLA class I histocompatibility antigen, B-67 alpha chain (MHC class I antigen B*67)	332.13	1,071	0.46416
Q29940	*HLA-B*	HLA class I histocompatibility antigen, B-59 alpha chain (MHC class I antigen B*59)	451.76	1,072	0.46419
Q31610	*HLA-B*	HLA class I histocompatibility antigen, B-81 alpha chain (B'DT) (MHC class I antigen B*81)	332.13	1,071	0.46416
Q31612	*HLA-B*	HLA class I histocompatibility antigen, B-73 alpha chain (MHC class I antigen B*73)	32,036.80	1,283	0.49117
Q95365	*HLA-B*	HLA class I histocompatibility antigen, B-38 alpha chain (Bw-4) (MHC class I antigen B*38)	345.40	1,073	0.46418
Q30154	*HLA-DRB5*	HLA class II histocompatibility antigen, DR beta 5 chain (DR beta-5) (DR2-beta-2) (Dw2) (MHC class II antigen DRB5)	26,778.34	1,290	0.48645
P09601	*HMOX1*	Heme oxygenase 1 (HO-1) (EC 1.14.14.18)	27,916.14	515	0.47913
P00738	*HP*	Haptoglobin (Zonulin) [Cleaved into: Haptoglobin alpha chain; Haptoglobin beta chain]	44,329.89	554	0.47580
P05362	*ICAM1*	Intercellular adhesion molecule 1 (ICAM-1) (Major group rhinovirus receptor) (CD antigen CD54)	68,177.95	1,690	0.50307
P01562	*IFNA1*	Interferon alpha-1/13 (IFN-alpha-1/13) (Interferon alpha-D) (LeIF D)	3,160.95	255	0.44405

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P22301	*IL10*	Interleukin-10 (IL-10) (Cytokine synthesis inhibitory factor) (CSIF)	60,689.17	1,027	0.49031
P35225	*IL13*	Interleukin-13 (IL-13)	16,060.33	589	0.46863
P01584	*IL1B*	Interleukin-1 beta (IL-1 beta) (Catabolin)	43,577.22	866	0.48631
P18510	*IL1RN*	Interleukin-1 receptor antagonist protein (IL-1RN) (IL-1ra) (IRAP) (ICIL-1RA) (IL1 inhibitor) (Anakinra)	4,992.78	340	0.45408
P05112	*IL4*	Interleukin-4 (IL-4) (B-cell stimulatory factor 1) (BSF-1) (Binetrakin) (Lymphocyte stimulatory factor 1) (Pitrakinra)	50,896.72	899	0.48677
O15327	*INPP4B*	Type II inositol 3,4-bisphosphate 4-phosphatase (EC 3.1.3.66) (Inositol polyphosphate 4-phosphatase type II)	6,224.22	263	0.45068
P10914	*IRF1*	Interferon regulatory factor 1 (IRF-1)	24,455.19	674	0.48029
Q96A59	*MARVELD3*	MARVEL domain-containing protein 3	2,526.89	84	0.39215
P11226	*MBL2*	Mannose-binding protein C (MBP-C) (Collectin-1) (MBP1) (Mannan-binding protein) (Mannose-binding lectin)	13,984.13	356	0.44608
P03971	*MIF*	Muellerian-inhibiting factor (Anti-Muellerian hormone) (AMH) (Muellerian-inhibiting substance) (MIS)	9,561.06	279	0.45369

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
P14174	*MIF*	Macrophage migration inhibitory factor (MIF) (EC 5.3.2.1) (Glycosylation-inhibiting factor) (GIF) (L-dopachrome isomerase) (L-dopachrome tautomerase) (EC 5.3.3.12) (Phenylpyruvate tautomerase)	4,797.51	241	0.45795
P11055	*MYH3*	Myosin-3 (Muscle embryonic myosin heavy chain) (Myosin heavy chain 3) (Myosin heavy chain, fast skeletal muscle, embryonic) (SMHCE)	26,135.36	1,122	0.49250
P60321	*NOS2*	Nanos homolog 2 (NOS-2)	3,112.42	123	0.42336
P35228	*NOS2, NOS2A*	Nitric oxide synthase, inducible (EC 1.14.13.39) (Hepatocyte NOS) (HEP-NOS) (Inducible NO synthase) (Inducible NOS) (iNOS) (NOS type II) (Peptidyl-cysteine S-nitrosylase NOS2)	14,718.65	425	0.47056
Q9H2C8	*OR51V1*	Olfactory receptor 51V1 (Odorant receptor HOR3'beta1) (Olfactory receptor 51A12) (Olfactory receptor OR11-36)	2,857.04	755	0.42798
P16284	*PECAM1*	Platelet endothelial cell adhesion molecule (PECAM-1) (EndoCAM) (GPIIA') (PECA1) (CD antigen CD31)	63,425.28	1,634	0.49978
P28065	*PSMB9*	Proteasome subunit beta type-9 (EC 3.4.25.1) (Low molecular mass protein 2) (Macropain chain 7) (Multicatalytic endopeptidase complex chain 7) (Proteasome chain 7) (Proteasome subunit beta-1i) (Really interesting new gene 12 protein)	13,964.38	417	0.46688

Continued on next page

Table 15 – continued from previous page

UniProt ID	Gene name	Description	Betweenness	Degree	Closeness
O75880	*SCO1*	Protein SCO1 homolog, mitochondrial	20,845.03	391	0.44647
O00206	*TLR4*	Toll-like receptor 4 (hToll) (CD antigen CD284)	111,490.44	1,360	0.50077
Q9NR96	*TLR9*	Toll-like receptor 9 (CD antigen CD289)	25,906.76	869	0.48774
P01375	*TNF*	Tumor necrosis factor (Cachectin) (TNF-alpha) (Tumor necrosis factor ligand superfamily member 2) (TNF-a) [Cleaved into: Tumor necrosis factor, membrane form (N-terminal fragment) (NTF); Intracellular domain 1 (ICD1); Intracellular domain 2 (ICD2); C-domain 1;	315,200.15	1,805	0.50908
Q13829	*TNFAIP1*	BTB/POZ domain-containing adapter for CUL3-mediated RhoA degradation protein 2 (hBACURD2) (BTB/POZ domain-containing protein TNFAIP1) (Protein B12) (Tumor necrosis tor, alpha-induced protein 1, endothelial)	7,341.54	278	0.44825
P0DSE2	*TRB*	M1-specific T cell receptor beta chain (TR beta chain TRBV19*01J2S7*01C*02)			
Q8NB14	*USP38*	Ubiquitin carboxyl-terminal hydrolase 38 (EC 3.4.19.12) (Deubiquitinating enzyme 38) (HP43.8KD) (Ubiquitin thioesterase 38) (Ubiquitin-specific-processing protease 38)	5,893.34	244	0.43786

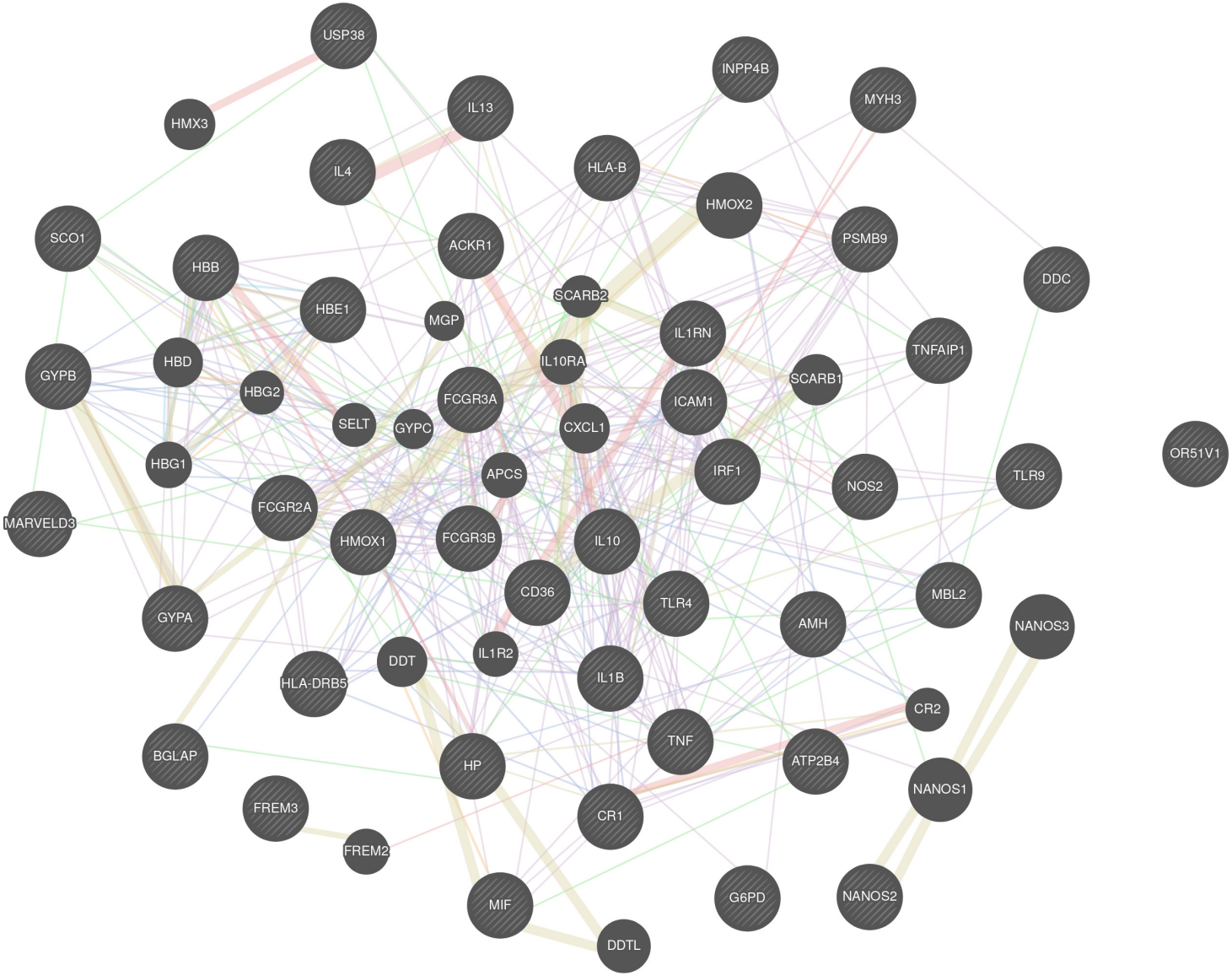

Figure 22. Functional interaction network between malaria–specific genes of the host and other host genes generated from genemania database. Genes are represented as nodes and interactions as edges.

4.7 Network Clustering

The ultimate aim of the GWAS behind identifying disease–associated genes is to map them onto the unified human functional network to identify hubs or clusters containing key genes or proteins regulating the structural integrity of the functional network. The centrality score for each node within the network was used to investigate key nodes or proteins in the functional network.
Based on the degree, betweenness and closeness metric, hubs formed within the network were categorized as either degree-based or structural hub. Degree-based hubs are formed by nodes with a degree score above the average degree score of approximately 406.
Similar to the criteria implemented in identifying key proteins with the parasite network, we filtered out nodes with betweenness and closeness score above the average for further structural and functional network analysis to elucidate key candidate proteins. These nodes are characterized with high degree score.

Also, we implemented the Blonde et al [113] clustering algorithm to cluster the human functional network. The disease-associated genes were mapped onto the network to identify subnetworks or communities containing candidate disease-associated gens and those encoding key proteins. These key proteins are considered to be more influential and significant in the network. The human functional network contained 760 key proteins within the 32 clusters. Out of the 32 clusters, 7 contained 78 malaria candidate genes whiles 2 of these 7 clusters contained 6 malaria-associated genes encoding key proteins. These key host malaria-associated genes (**Table 17**) could serve as essential targets of protective immunity against malaria particularly in relation to antibody–based therapies [270]. **Table 16** describes the distribution of proteins in the 7 clusters.

Table 16. Classification of the functional human network into clusters.

Cluster ID	Number of proteins	Number of key proteins	Number of candidate proteins	Number of candidate key proteins
0	7,620	136	4	0
1	7,12	24	1	0
2	3,964	186	28	5
3	2,107	46	5	0
4	930	120	1	0
5	1,302	113	38	1
6	1,218	49	0	0
Total	17,853	674	78	6

Table 17. Key malaria-associated genes found in the human functional network with their betweenness, degree, and closeness network centrality measures.

UniProt-ID	Gene name	Betweenness	Degree	Closeness
P22301	*IL10*	60,689.17	1,027	0.49031
P05362	*ICAM1*	68,177.95	1,690	0.50307
P01375	*TNF*	315,200.15	1,805	0.50908
P30480	*HLA-B*	76,201.46	1,330	0.49009
P16284	*PECAM1*	63,425.28	1,634	0.49978
O00206	*TLR4*	111,490.44	1,360	0.50077

4.8 Functional Analysis of Human Disease-Associated Candidate Genes Encoding Key Proteins

In this section, we investigate the molecular and biological functions of the candidate genes encoding candidate key proteins in the functional network using functional genomics databases (**Table 4**) and statistical methods such as the Bonferroni multiple test correction to estimate adjusted p-values. The p-values were estimated primarily by using the frequency of occurrence of each process in relation to the candidate genes. These candidate genes are involved in essential biological processes (**Table 18**) and pathways (**Table 19**) that are linked to the proper functioning of the biological system. Having established that nodes within a cluster might be involved in the same biological process, it is therefore possible that these key proteins within the clusters contribute significantly to similar processes [97]. We used the six key candidate proteins (**Table 17**) as target sets in order to perform disease-associated gene annotation to investigate statistically significant biological processes that influences malaria pathogenesis in human. 23 significantly enriched malaria-related biological processes described in **Table 18** were identified. These gene ontology groups comprised of those involved in cell immune and inflammatory response, regulation and production of transcription factors, biosynthetic processes, cell-cell adhesion, cell signaling and cell apoptotic processes.

GO:0042346 process responsible for regulation of NF-kappaB importation have been studied to be involved in immune and inflammatory responses particularly in eukaryotic cells. Down or negative regulation of NF-kappaB has been reported to be associated with *Plasmodium falciparum*-modulated endothelium transcriptome contributing to cerebral malaria [271]. GO:0045348 process responsible for positive regulation of major histocompatibility complex (MHC) class II biosynthetic process regulates immune response to malaria [272]. Pre-erythrocytic immunity to malaria (cerebral malaria) in Africa is linked to MHC antigens such that variations in class I and class II in these antigens contribute significantly to malaria susceptibility thus, reduced or increased host immune response [272]. Also, other processes such as GO:0032689, GO:0032715, GO:0002740 and GO:0032729 serves as immunological mediating processes that influence malaria susceptibility by either conferring protection or influencing disease pathogenesis.

Activation and regulation of NLRP3 inflammasomes, immune system receptors, controls the activation of caspase-1 and induce inflammation in response to infectious pathogens [273]. Due to their influence on a wide range of diseases, their dysfunctioning results in the initiation or progression of diseases.

Endothelial cell apoptosis has been studied to contribute to malaria severity. For instance, heme-induced microvasculature endothelial cell apoptosis mediated by proinflammatory and proapoptotic

pathways contributes significantly to severe malaria.

Table 18. Statistically significant biological processes of key human malaria-associated genes. The GO term level is a numerical representation of the biological process from the root (level 0) of the GO hierarchical structure.

Gene Ontology (GO)-ID Process	Gene Ontology Term Name	GO Term Level	p-value	Adjusted p-value
GO:0042346	positive regulation of NF-kappaB import into nucleus	12	2.00161e-05	0.00432
GO:0045348	positive regulation of MHC class II biosynthetic process	7	2.40637e-06	0.00052
GO:0032689	negative regulation of interferon-gamma production	6	3.10034e-05	0.00670
GO:0007157	heterophilic cell-cell adhesion via plasma membrane cell adhesion molecules	5	0.00012	0.02714
GO:2000352	negative regulation of endothelial cell apoptotic process	9	2.70760e-05	0.00585
GO:0032715	negative regulation of interleukin-6 production	6	8.99764e-08	1.9434e-05
GO:2000343	positive regulation of chemokine (C-X-C motif) ligand	7	1.68841e-05	0.00364
GO:0032729	positive regulation of interferon-gamma production	6	0.00012	0.02713
GO:0070374	positive regulation of ERK1 and ERK2 cascade	11	2.8883e-05	0.00623
GO:0050830	defense response to Gram-positive bacterium	7	2.65221e-05	0.00572
GO:0034116	positive regulation of heterotypic cell-cell adhesion	6	1.68841e-05	0.00364
GO:0044130	negative regulation of growth of symbiont in host	7	6.07930e-06	0.00131
GO:0030198	extracellular matrix organization	4	5.39819e-05	0.01166

Continued on next page

Table 18 – continued from previous page

Gene Ontology (GO)-ID Process	Gene Ontology Term Name	GO Term Level	p-value	Adjusted p-value
GO:0045416	positive regulation of interleukin-8 biosynthetic process	8	2.40637e-06	0.00051
GO:0032755	positive regulation of interleukin-6 production	6	0.00016	0.03562
GO:0002740	negative regulation of cytokine secretion involved in immune response	10	1.00303e-06	0.00021
GO:0045429	positive regulation of nitric oxide biosynthetic process	8	1.23374e-07	2.665e-05
GO:0043032	positive regulation of macrophage activation	7	1.02165e-05	0.00220
GO:1904999	positive regulation of leukocyte adhesion to arterial endothelial cell	8	2.00663e-07	4.3343e-05
GO:0031663	lipopolysaccharide-mediated signaling pathway	9	2.42641e-08	5.2410e-06
GO:1904707	positive regulation of vascular smooth muscle cell proliferation	7	0.00016	0.03663
GO:0032800	receptor biosynthetic process	6	6.68766e-07	0.00014
GO:1900227	positive regulation of NLRP3 inflammasome complex assembly	7	2.70759e-05	0.00584

4.9 Enrichment Analysis

In this section, we performed pathway enrichment analysis to investigate statistically significant pathways inferred from KEGG database and UniProtKB-GOA data set that potentially play essential roles in malaria pathogenesis. Similar to investigating the biological processes, we implemented a Bonferroni multiple test correction adjusted hypergeometric test to identify enriched pathways linked to the candidate genes. Also, we estimated the p-values for these pathways using their frequency of occurrence in relation to the key proteins and the human proteome in general. **Table 19** describes pathways which potentially contribute to malaria pathogenesis. These pathways are involved in immune response, cell-cell signaling and production of key transcription factors specific to disease pathogenesis. Interestingly, artemisinin derivatives, particularly artesunate mostly used in SSA are involved in most of the processes and pathways identified in our study. For instance, it blocks the production of *IL-1β, IL-6* and *IL-8* [274]. It also inhibits phosphoinositide 3-kinase (PI3K)/Akt signaling pathway and lipopolysaccharide-induced production of TNF-α, IL-6 and nitric oxide (NO) [274]. Also, artemisinin

derivatives are involved in NF-kappaB transcription activities and anti-inflammatory processes. Among the pathways are the *Malaria pathway* (hsa05144), *Natural killer cell mediated cytotoxicity pathway* and other disease-enriched pathways including but not limited to Tuberculosis, Autoimmune thyroid disease, Hematopoietic cell lineage and Type I diabetes mellitus.

Table 19. Statistically Significant Pathways of Human Key Malaria-Associated Genes.

textbfKEGG-Pathway ID	**KEGG-Pathway-Name**	**p-value**	**Adjusted p-value**
path:hsa05133	Pertussis	8.77858e-06	0.00107
path:hsa04940	Type I diabetes mellitus	9.00124e-05	0.01098
path:hsa05144	Malaria	0.0	0.0
path:hsa05310	Asthma	6.52960e-07	7.966e-05
path:hsa04145	Phagosome	1.46003e-06	0.00017
path:hsa05146	Amoebiasis	0.00014	0.01745
path:hsa04640	Hematopoietic cell lineage	1.70661e-06	0.00020
path:hsa05330	Allograft rejection	3.03329e-06	0.00037
path:hsa05162	Measles	8.93218e-05	0.0108
path:hsa04650	Natural killer cell mediated cytotoxicity	0.00014	0.01724
path:hsa04657	IL - 17 signaling pathway	0.00037	0.04624
path:hsa05152	Tuberculosis	1.92895e-10	2.35332e-08
path:hsa05150	Staphylococcus aureus infection	4.40440e-08	5.37336e-06
path:hsa05142	Chagas disease (American trypanosomiasis)	7.77032e-05	0.00947
path:hsa05143	African trypanosomiasis	7.32790e-10	8.9400e-08
path:hsa05140	Leishmaniasis	1.16192e-11	1.41754e-09
path:hsa05321	Inflammatory bowel disease (IBD)	9.68744e-08	1.18186e-05
path:hsa05322	Systemic lupus erythematosus	3.51918e-05	0.00429
path:hsa05323	Rheumatoid arthritis	0.00027	0.03331
path:hsa05320	Autoimmune thyroid disease	9.62632e-06	0.00117
path:hsa05332	Graft - versus - host disease	0.00010	0.01229

4.10 Summary

In this chapter, we constructed human functional network from human proteome and protein-protein functional interaction data retrieved from literature and databases. We also identified human malaria associated genes by mapping significant SNPs from GWAS summary statistics data among Gambia, Malawi and Kenya population to gene level. We extracted host malaria–specific network from the generated human network. These 79 identified genes were mapped onto the functional network to identify clusters and disease-associated candidate genes encoding key proteins in the network. We identified 32 clusters of which 7 contained candidate proteins. Out of the 7 clusters were 2 clusters containing 6 candidate key proteins (P22301 (*IL10*), P05362 (*ICAM1*), P01375 (*TNF*), P30480 (*HLA-B*), P16284 (*PECAM1*) and O00206 (*TLR4*)) highly relevant to malaria pathogenesis. We performed disease-associated gene annotation enrichment analysis on the candidate key proteins to elucidate statistically significant processes and pathways related to the disease.

CHAPTER 5

5 Combined Human–*P. falciparum* Proteome Functional Networks

The manifestation of malaria depends on the action of host system on the pathogen and the reaction from the pathogen to the human host. From the parasite's perspective, these actions mainly includes the pathogen's ability to invade the host system as well as evading host immune response by affecting essential pathways and other processes in the host [275]. These interactions are key to the survival of the pathogen within the host. In contrast to the parasite optimal survival mechanisms, the host response to the invader through initiation of signaling cascade to increase defense mechanism. The purpose of constructing host-pathogen network is to understand the interplay between the inter-species PPIs to elucidate critical overlapping nodes underlying functional interactions and mechanisms essential for disease aetiology. Also, host–pathogen network analysis would contribute to understanding possible interactions and mechanisms underlying the development of resistance to current drugs. In this chapter, we focused mainly on interactions between our identified putative drug targets (**Table 20**) and the host. Also, we investigate functional interactions between malaria drug resistant genes (**Table 2**) and their interactions with the identified host malaria–susceptible genes (**Table 15**) and its contribution to adaptation to host immunity and development of resistance.

5.1 Functional Interaction Datasets for Constructing Host–Pathogen Functional Network

To construct the host-pathogen network, the generated host and pathogen network together with other host-pathogen interaction dataset retrieved from literature and databases such as InterPro, BioGrid, DIP, HPIDB and MINT described in **Table 4** were used together. Also, sequence BLAST between the host and pathogen were used in this section to develop the unified host-pathogen network. We applied scoring algorithms implemented in **Chapters 3 and 4** to score the functional datasets prior to generating the network. The interaction score between each pair-wise proteins ranged from 0.3 to 1.

5.2 Construction of Human–*P. falciparum* Functional Network

The defined scheme implemented in generating separate organism network was applied in this section. The unified network comprised of 31,512 host-pathogen functional interactions between 8,023 nodes. However, as most of the drug resistance genes were not yet reviewed during the time of this study, we extracted interactions between the drug resistant genes and the host genes from interpro and sequence blast. In total, 68,285 interactions and 8,690 nodes were used for further downstream analysis. The combined confidence score between pair–wise proteins (p and q) in the network is given as shown in equation 12.

$$S_{pq} = 1 - \prod_{t=1}^{9}(1 - s_{pq}^{d}) \tag{12}$$

where s_{pq}^{d} is defined as the confidence score between p and q proteins using data type d.

Figure 23 shows the distribution of the node degrees following the power law distribution. The graphs shows that many nodes within the network are characterized by less degrees score whereas the converse is true for few nodes.

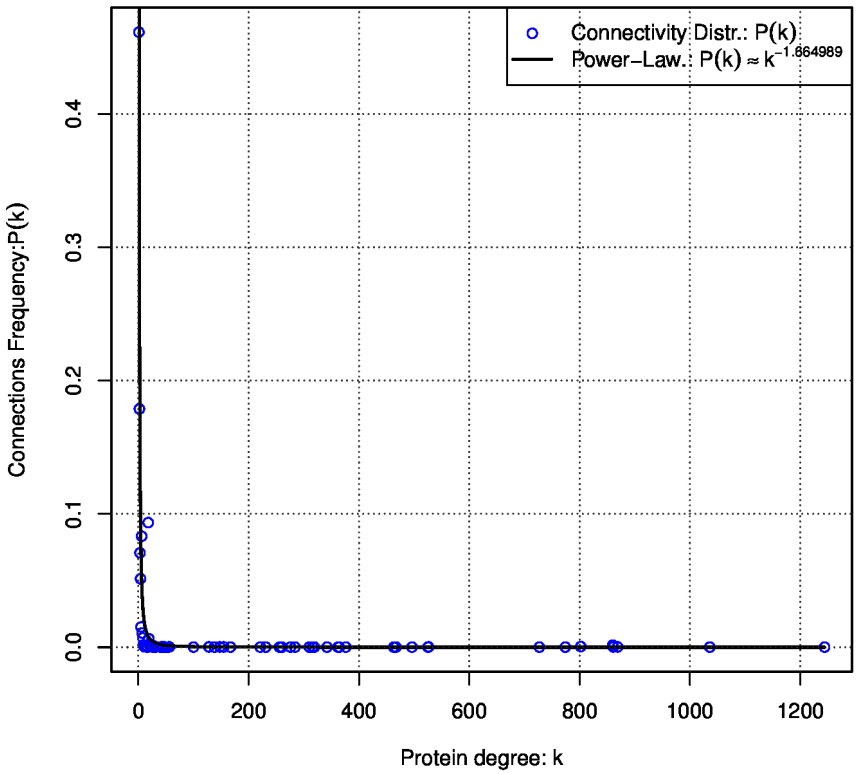

Figure 23. Power law degree distribution of nodes in the unified human-*falciparum* functional network. The distribution shows that the network is made up of fewer nodes with higher degree.

5.3 Investigating Potential Host–Pathogen Interactions Influencing Resistance to Artemisinin

To explore potential mechanisms that could likely account for reduced artemisinin sensitivity phenotype, susceptibility and resistance development within the African populations with a view of controlling resistance, we investigated interactions between artemisinin–resistant genes and the host malaria susceptible genes identified from the GWAS data. We started by filtering the host–pathogen network **(Chapter 5 section 5.2)** for pairwise interactions that contained either *pfk13*, *pfmdr1*, *pfcrt*, *pfdhps*, *dhfr* or any of the host malaria susceptible proteins(**Table 15**). We identified 410 functional interactions between 431 nodes. We assessed the network formed from the extracted interactions to identify subnetworks.

We observed disjoint subnetworks (**Figure 24**) between host susceptible genes and the parasite resistant genes formed within the network. We observed that *pfk13* protein (*PF3D7_1343700*) has direct functional interactions (**Figure 25**) with other host kelch–like proteins highly expressed in the host. Among the Kelch–like proteins are Kelch-like protein 2 (*KLHL2*), Kelch-like protein 18 (*O94889*), Kelch repeat and BTB domain-containing protein 8 (*KBTBD8*), Kelch domain-containing protein 8A (*KLHDC8A*), Kelch repeat and BTB domain-containing protein 12 (*KBTBD12*), Kelch-like protein 30 (*KLHL30*) and Kelch-like protein 13 (*KLHL13*) involved mainly in protein ubiquitination, translation regulation, post–translational protein modification and regulation of cytokinesis [93]. These interactions suggests that, regulations of protein–associated processes and pathways play a critical role in the biological activities of artemisinin particularly protein ubiquitination, knowing that the drug targets various proteins in the parasite [276].

Interestingly, the *pfk13* protein interacts with other host regulatory genes involved in essential processes such as transcription regulation, cell–surface, cell–cell signaling and regulation of phosphorylation [93]. Among the regulatory genes include the Transcriptional regulator Kaiso (*ZBTB33*), Zinc finger and BTB domain-containing protein 17 (*ZBTB17*), BTB/POZ domain-containing protein 10 (*Q9H3F6*), Zinc finger and BTB domain-containing protein 10 (*ZBTB10*), Myoneurin (*MYNN*), Nucleoprotein TPR (*TPR*) and Gigaxonin (*GAN*).

In our analysis, we did not observe a direct functional interaction between *pfk13* and the host malaria–associated genes. However, we observed a subnetwork (blue nodes) centralized by Coagulation factor XIII A chain(*F13A1*) as shown in **Figure 24**. From our data, these interactions could imply that, the contribution of *pfk13* to artemisinin resistance development is likely to be modulated by other latent genes that can provide more insight.

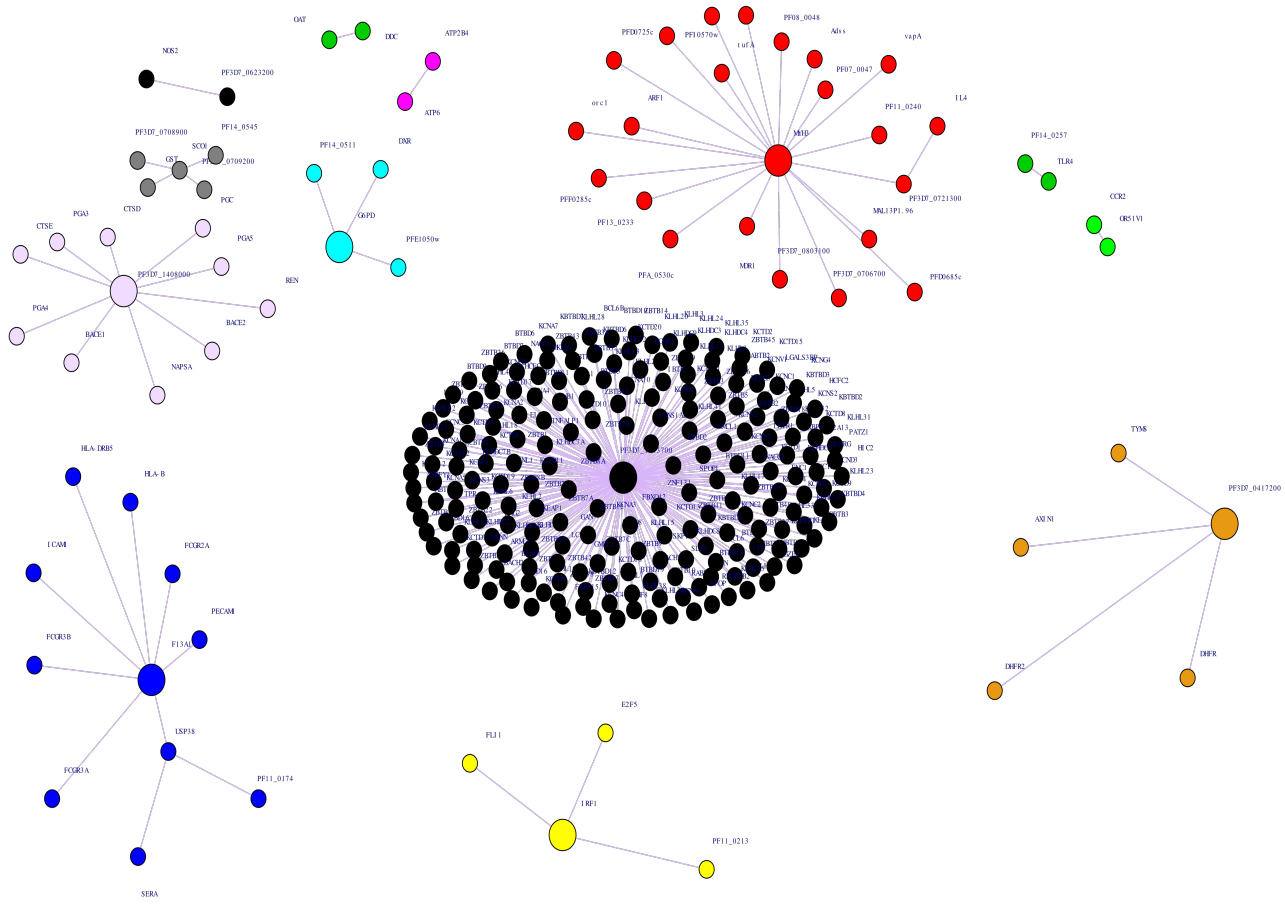

Figure 24. Subnetworks formed between malaria selective genes and host malaria susceptible genes.

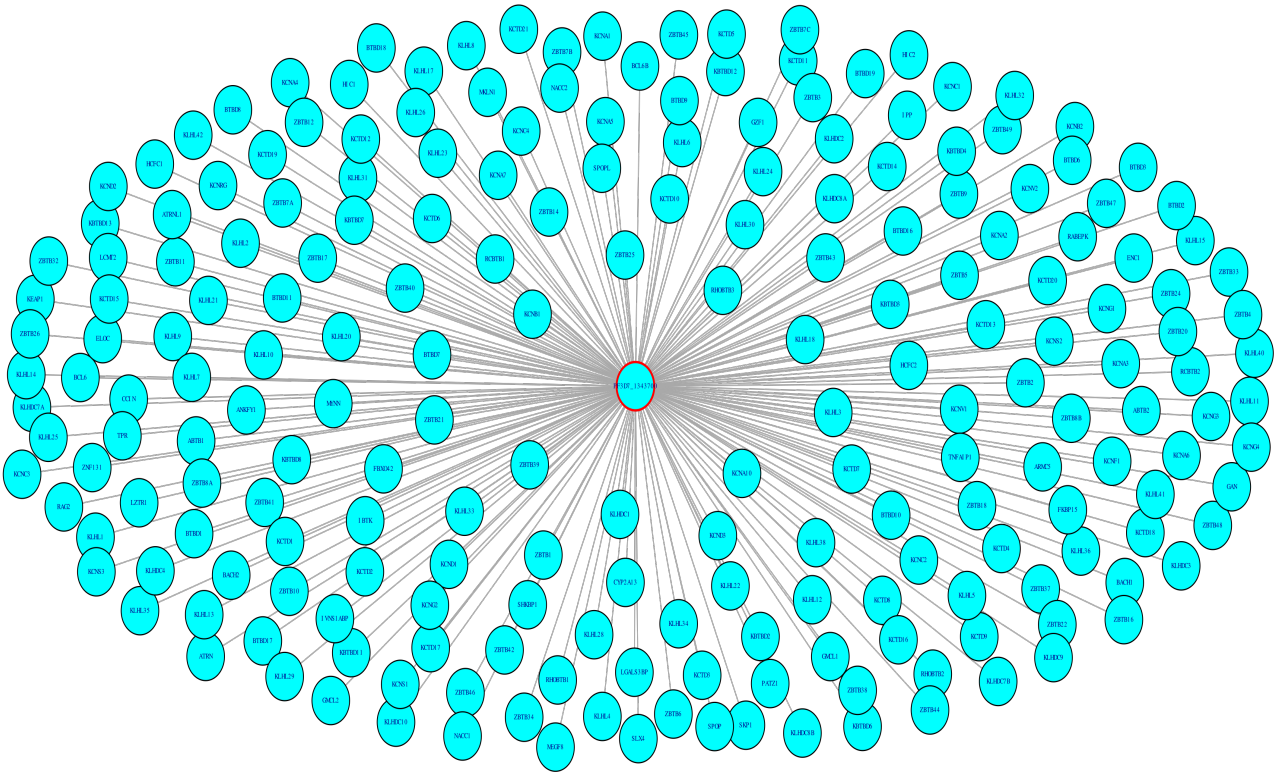

Figure 25. Subnetwork comprising of 295 interactions formed between *pfk13*(central node) within the host–pathogen network.

5.4 Investigating Potential Host–Pathogen Interactions Influencing Drug Resistance and Host Immune Tolerance through C6KTD2 and C6KTB7

Within the host–pathogen functional network, 6,139 human proteins functionally interacts with 108 candidate pathogen proteins. We investigated the functional interactions between our key candidate proteins (**Table 20**) and host proteins within the constructed host–pathogen network. The purpose was to investigate whether these key candidate proteins influence the structural integrity and the flow of information within the host-pathogen generated functional network.
First, we analyzed the nodes that are common to subnetworks formed by C6KTD2 (**Figure 26**) and C6KTB7 (**Figure 27**) to elucidate the connectivity between these clusters within the human–pathogen functional network. Host proteins with uniprot ID's Q99728, Q86YT6, Q96KQ7, Q9H9B1, Q7L622, Q96AX9, Q9P2G1 and Q9P2R3 were found to functionally connect these clusters as shown in **Figure 28**. These intersecting proteins (yellow nodes in **Figure 28**) are involved in protein ubiquitination, positive regulation of cell apoptotic process, signal transduction, regulatory processes and histone methylation [93].

Knowing that genes and their products such as proteins interact with the cellular environment in multiple levels to ensure proper functioning of the cell and biological pathways [252], our second analysis in this section focused on elucidating genes or proteins and their related pathways that could possibly interfere with host immune response and contribute to resistance under drug pressure. This analysis is highly essential in drug research and proper understanding of the pathogen's behaviour. Such analysis would help to understand the modes and dynamic patterns underlying such processes. Drug resistance machinery remains to be the natural driving force influencing the parasites survival when exposed to antimalarials. Usually, the resistant mechanism and pathogen adaptiveness to host response involves different pathways compared to the paths resulting to target inhibition by a drug and effective immune-related pathways [277].

In this section, we analyzed the host-pathogen functional network to investigate shortest paths between the key candidate proteins identified in the host and pathogen functional network to gain insight on the possible routes for innate immune response interference and drug resistance development. Studies have shown that, shortest path analysis of a functional network yields high coverage compared to direct neighbours within the network [277]. Shortest paths between host-pathogen disease associated candidate key genes herein refers to the minimum number of edges required to connect these genes.

Longer paths consist of more nodes involved in a cascade of signaling process to trigger innate immune response by inducing the production of chemokines and cytokines upon parasite infection. It is therefore a measure of information relay between the candidate key genes, thus, the shorter the path the quicker the transmission. It is of noteworthy that, shortest path lengths between the pathogen disease-associated genes and human disease-associated genes conferring immunity in the functional network are the most feasible routes of drug resistance development [277]. Due to the dynamic pattern of the network, shortest distance highly increases the susceptibility of human to malaria infection.

The shortest paths identified and their associated pathways as shown in **Figures 29** and **30** suggests that inhibition or alteration to the proper function of each path might help the parasites to survive immune responses, thus, aggregation of small effects. The development of adaptive immunity and drug resistance is expected to happen when the parasite undergoes diversity throughout time such that they evade the host system when they become tolerant and establish different mechanisms to interfere the host's response. Parasite diversity could be in the form of the modulation of surface proteins that results in the development of antiparasitic immunity leading to severe malaria. These interferences can also be in the form of production of effector mechanisms that can down-regulate innate immunity. We observed that, the dynamic pattern to resistance and immune adaptiveness is mediated by other human-specific genes or proteins conferring immunity herein referred to as co-targets. The co-targets are very critical in the resistant network such that they strongly influence the resistance machinery [277].

In our study, we identified human immune-related genes and pathways that could be inhibited by the pathogen, knowing that the pathology of malaria is immune mediated.

We observed potential resistance pathways between host malaria-associated protein O00206 (Toll-like receptor 4, *TLR4*) and pathogen proteins C6KTB7 (Putative E3 ubiquitin-protein ligase protein PFF1365c) and C6KTD2 (Putative histone-lysine N-methyltransferase 1, *SET1*).

Severe malaria is associated with an increased level of pro-inflammatory cytokines such as interleukin *(IL)-12*, *IL-8*, and interferon (*IFN*)-γ in the affected person which helps to modulate defense against the infection [278]. This is because, the severity of malaria is proportional to the flawlessness in inflammatory response by the host.

TLR4, a pathogen–recognition receptor, detects pathogen-associated molecular mechanisms in the body and initiates immune response through activation of signaling cascades such as nuclear factor-kB, mitogen-activated protein kinase (MAPK) and *Plasmodium* antigens [278]. Toll-like receptor 4, *TLR4* (O00206), and its immune-related signaling pathways have been reported to contribute significantly to *Plasmodium falciparum* growth and malaria pathogenesis, such that dysregulation and dysfunction of the gene increases malaria severity, symptomatic malaria, severe malarial anemia and resistance in Africa [279]. This implies that inhibition of such related pathways by C6KTB7 and C6KTD2 will contribute significantly to resistance.

In our analysis, we identified some mediators (nodes) connecting the targets (**Figures 28**). These mediators are host proteins involved in various immune signaling pathways. These proteins are involved in processes such as the *toll–like receptor signaling pathway*, *tumor necrosis factor–mediated signaling pathway*, *NF-kappaB signaling pathway*, *inflammatory signaling pathway*, adherence–mediated pathways, cell apoptosis pathway and the regulation of T-cell cytokine.

Furthermore, we investigated possible functional interactions or shortest paths between key proteins in human and the pathogen that could lead to resistance. We identified O00206 (Toll-like receptor

4, *TLR4*) human candidate key protein to interact with C6KTB7 (Putative E3 ubiquitin-protein ligase protein PFF1365c) 15 possible paths through mediating nodes which are specific to human. From our analysis, we propose the essential pathways of the mediating nodes to contribute to host immune adaptiveness and drug resistance development. **Table 21** shows the functional interactions between O00206 and C6KTB7.

Table 20. Degree, betweenness and closeness centrality score of C6KTD2 and C6KTB7 within the parasite unified functional network.

Uniprot ID	Gene name	Description	Betweenness	Degree	Closeness
C6KTD2	*SET1*	Putative histone-lysine N-methyltransferase 1	1634413.73	525	0.30214
C6KTB7	*PFF1365c*	Putative E3 ubiquitin-protein ligase protein PFF1365c	1169508.41	284	0.26346

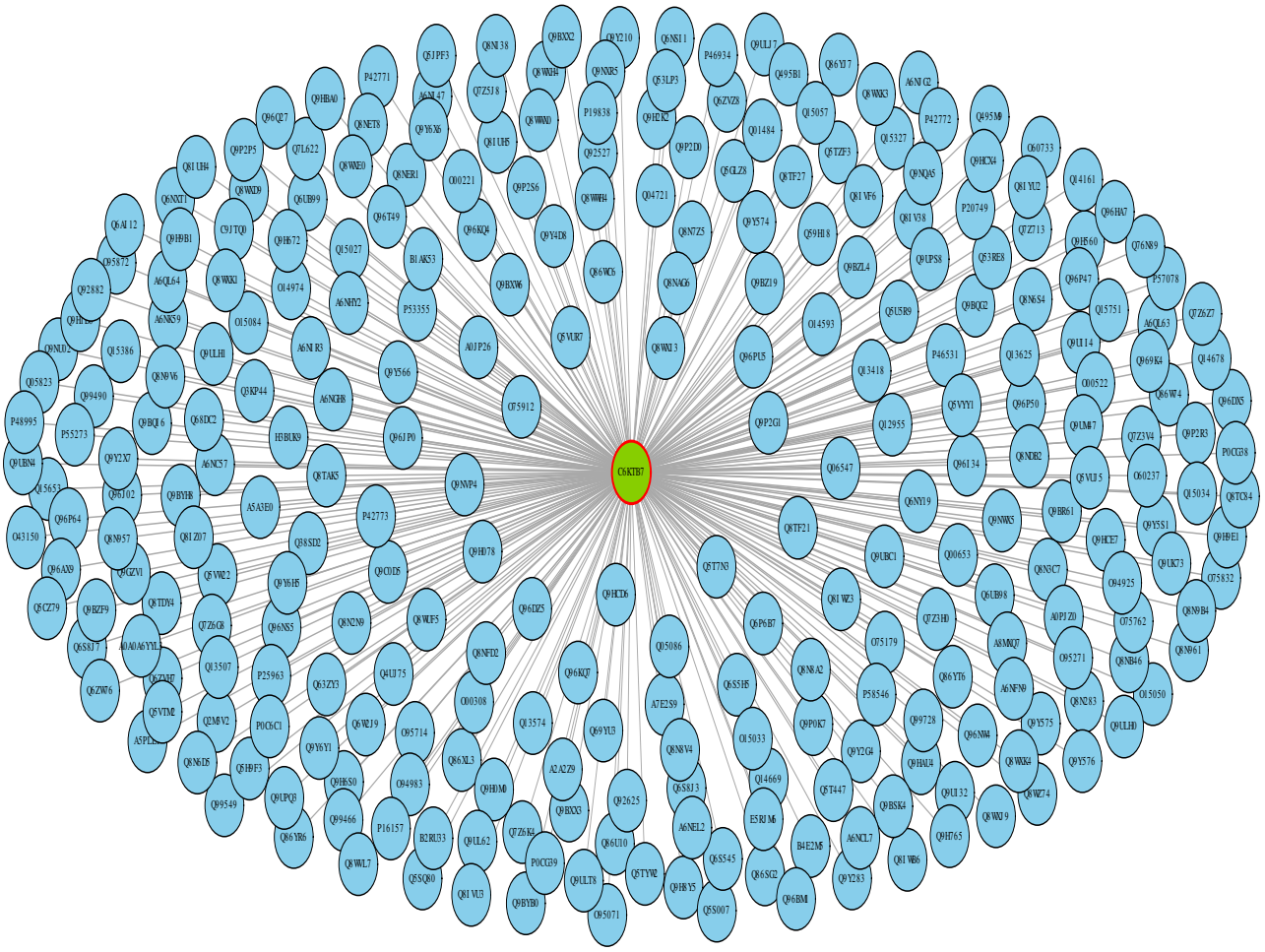

Figure 26. Functional interactions between C6KTB7 (green node) and other nodes in the unified human-pathogen functional network. C6KTB7 functionally interacts with 284 human proteins (skyblue nodes) in the unified host-pathogen network.

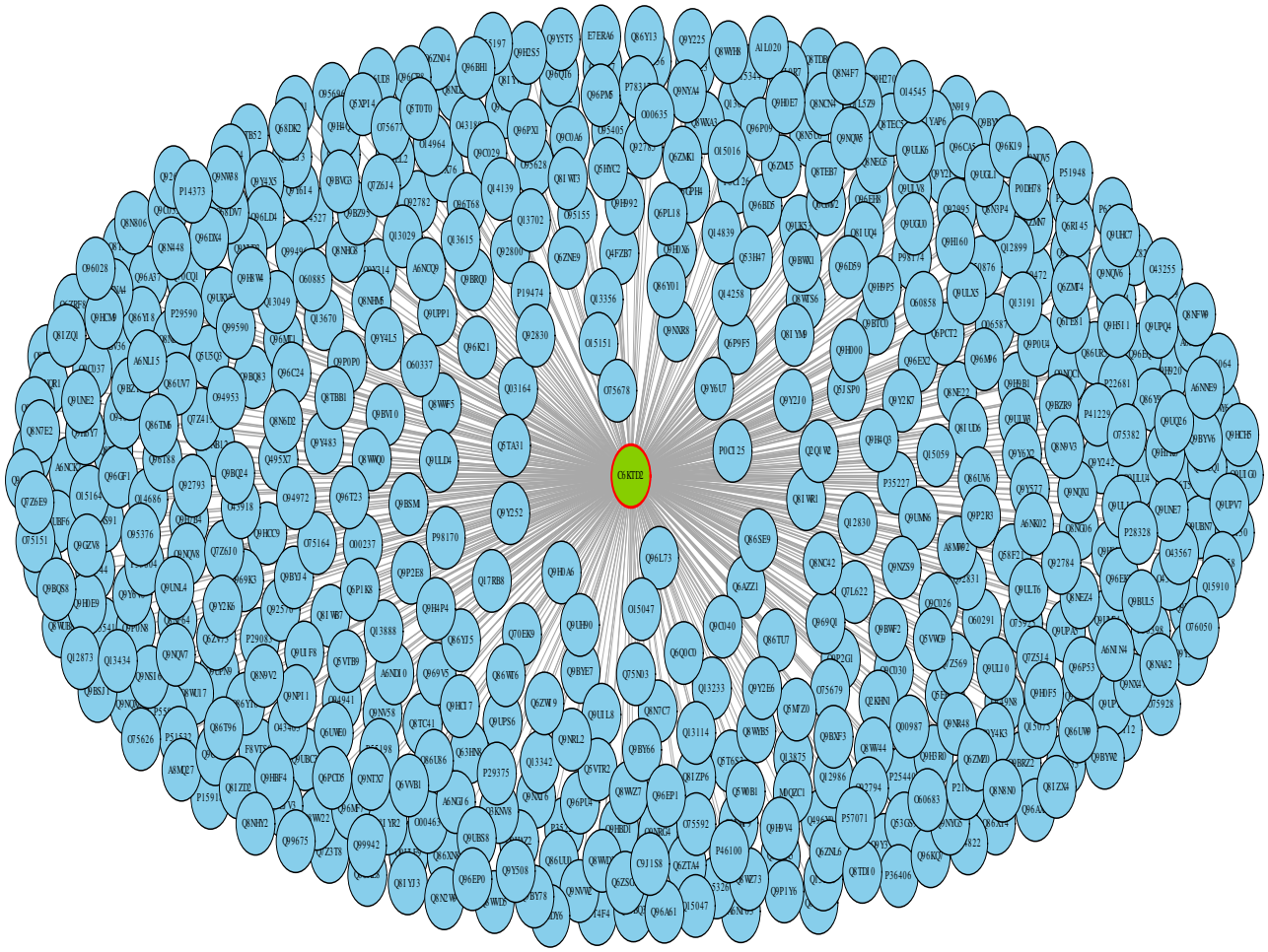

Figure 27. Functional interactions between C6KTD2 (green node) and human proteins in the unified human-pathogen functional network. C6KTD2 interacts with 525 human proteins (skyblue nodes).

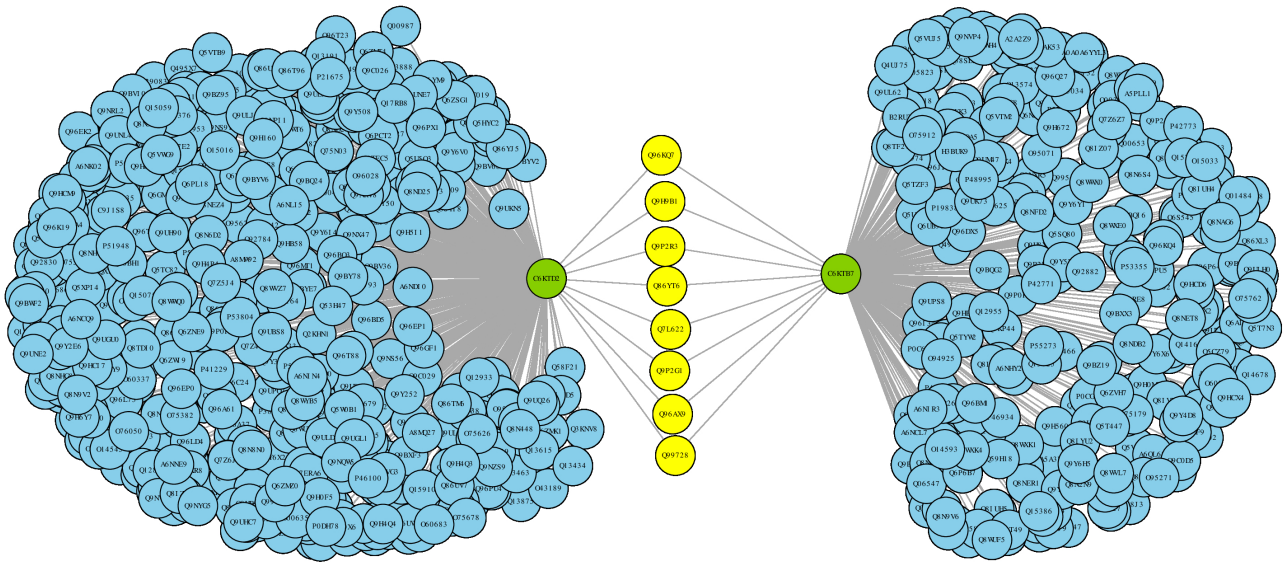

Figure 28. Investigating the shared proteins (yellow nodes) that connects clusters formed by C6KTD2(left green node) and C6KTB7(right green node) in the unified human-pathogen functional network.

From the shortest paths networks described in **Figures 29** and **30**, C6KTD2 has relatively higher short paths as compared to C6KTB7.

Table 21. Shortest paths linking O00206 and C6KTB7 nodes within the host-pathogen unified functional network.

Human protein	Mediator	Parasite protein	Potential inhibitory process
O00206	Q9BYH8	C6KTB7	T cell receptor signaling pathway
O00206	Q05823	C6KTB7	Interferon alpha/beta signaling
O00206	Q5S007	C6KTB7	canonical Wnt signaling pathway
O00206	Q38SD2	C6KTB7	canonical Wnt signaling pathway
O00206	O75762	C6KTB7	cell surface receptor signaling pathway
O00206	Q96HA7	C6KTB7	cytoplasmic sequestering of transcription factor
O00206	Q00653	C6KTB7	NIK/NF-kappaB signaling
O00206	P25963	C6KTB7	I-kappaB kinase/NF-kappaB signaling
O00206	P46531	C6KTB7	immune response
O00206	P20749	C6KTB7	antimicrobial humoral response
O00206	P42771	C6KTB7	regulation of NF-kappaB transcription factor activity
O00206	Q8NI38	C6KTB7	inflammatory response
O00206	P19838	C6KTB7	apoptotic process
O00206	Q8NDB2	C6KTB7	B cell activation
O00206	Q15653	C6KTB7	signal transduction

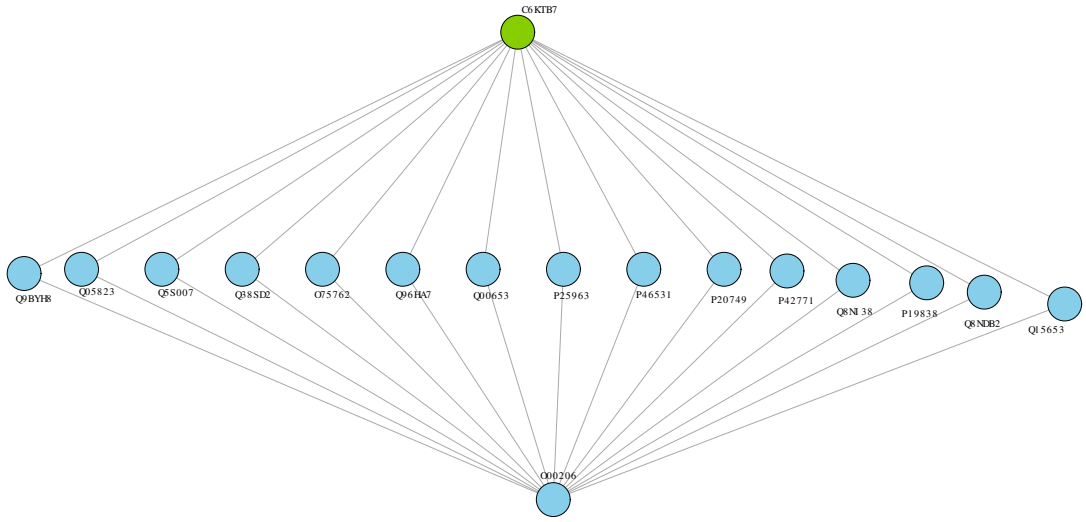

Figure 29. Predicted functional network that could influence resistance and host adaptiveness between C6KTB7 (green node) and O00206 (bottom skyblue node) via co–targets (central skyblue nodes) in the host-pathogen network.

Table 22. Shortest paths linking O00206 and C6KTD2 within the host-pathogen unified functional network.

Human protein	Mediator	Parasite protein	Potential inhibitory process
O00206	Q13114	C6KTD2	apoptotic process
O00206	Q86WT6	C6KTD2	protein ubiquitination pathway
O00206	Q12933	C6KTD2	protein ubiquitination pathway
O00206	Q9UPN9	C6KTD2	protein ubiquitination pathway
O00206	P55895	C6KTD2	MAPK6/MAPK4 signaling
O00206	O75626	C6KTD2	adaptive immune response
O00206	Q96CA5	C6KTD2	protein ubiquitination
O00206	Q8NHM5	C6KTD2	positive regulation of cell growth
O00206	Q9BUZ4	C6KTD2	activation of NF-kappaB-inducing kinase activity
O00206	Q96P53	C6KTD2	positive regulation of protein phosphorylation
O00206	P15918	C6KTD2	adaptive immune response
O00206	P19474	C6KTD2	innate immune response
O00206	Q9Y4K3	C6KTD2	activation of MAPK activity
O00206	Q6PCT2	C6KTD2	post-translational protein modification
O00206	Q14258	C6KTD2	innate immune response

Continued on next page

Table 22 – continued from previous page

Human protein	Mediator	Parasite protein	Potential inhibitory process
O00206	Q9UNE7	C6KTD2	protein ubiquitination pathway
O00206	Q15075	C6KTD2	endocytosis
O00206	Q8IWB7	C6KTD2	positive regulation of toll-like receptor 3 and 4 signaling pathway
O00206	Q09472	C6KTD2	apoptotic process
O00206	Q8IYM9	C6KTD2	interferon-gamma-mediated signaling pathway
O00206	Q9Y2K7	C6KTD2	ubiquitin conjugation pathway
O00206	Q9NQV6	C6KTD2	regulation of gene expression
O00206	Q92793	C6KTD2	positive regulation of type I interferon production
O00206	Q6UWE0	C6KTD2	protein ubiquitination pathway
O00206	Q8WVD3	C6KTD2	protein ubiquitination pathway
O00206	Q8TCQ1	C6KTD2	protein ubiquitination pathway
O00206	Q6Q0C0	C6KTD2	protein ubiquitination pathway
O00206	O00463	C6KTD2	apoptotic process
O00206	Q13233	C6KTD2	protein phosphorylation
O00206	P98170	C6KTD2	regulation of innate immune response
O00206	Q6ZMZ0	C6KTD2	protein ubiquitination pathway
O00206	O14964	C6KTD2	protein ubiquitination pathway
O00206	Q9NWF9	C6KTD2	protein ubiquitination pathway

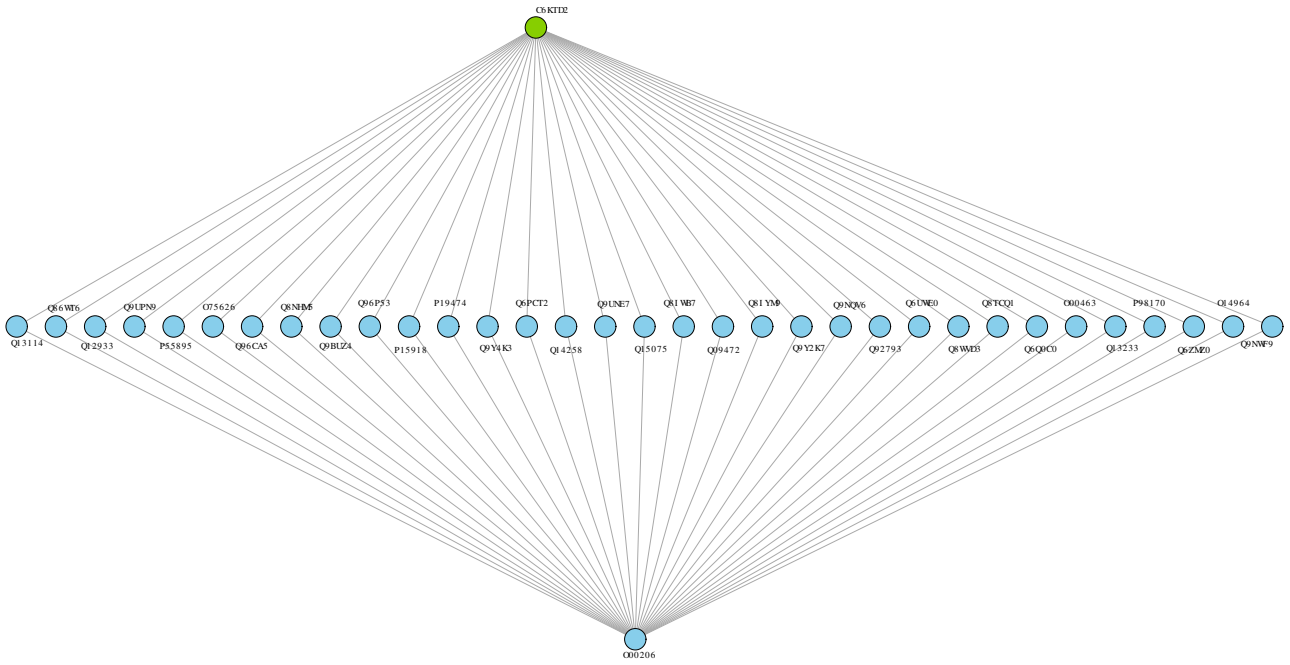

Figure 30. Shortest possible resistance pathways between *C6KTD2* (green node) and *O00206* (skyblue node) via co–targets (central skyblue nodes) in the host-pathogen network.

5.5 Summary

In this chapter, we focused on two main analysis to assemble a unified human-*Plasmodium falciparum* functional network from host and pathogen proteome and other host-pathogen interacting data retrieved from databases and literature. In the first analysis, we focused on investigating functional interactions between parasite resistant genes and their interactions with host malaria susceptible genes to explore potential mechanisms that could account for drug resistance during drug pressure and parasite diversity. Our results showed that, *pfk13* forms a subnetwork (**Figure 25**) with other essential regulatory host proteins but not with the host malaria susceptible genes. We propose that as the parasite undergoes diversity, the surface protein will evolve and adapt to essential processes contributing to the production of cytokines regulating immunity and proper functioning of the system. During the process of evolving under the influence of drug pressure, there is a likelihood of blockage in the inhibition and stimulatory effect of artemisinin on the host such that the host defense system such as antibody production and T cell response becomes overwhelmed and suppressed by the pathogen's activities thus resulting in increased parasite clearance time and drug resistance.

In the second analysis, we investigated the network focusing on the functional interactions between our identified key candidate proteins. We predicted possible resistant pathways and adaptive immune pathways by evaluating shortest paths between the host and pathogen key candidate proteins. We observed that, the resistance machinery associated to these targets may arise from multiple pathways.

CHAPTER 6

6 Predicting Repurposable Drugs for Malaria Treatment Based on Implicit Semantic Similarity

The development of human disease ontology terms [280] have provided an enriched platform of human disease data to evaluate similarities between various diseases of different disorder class based on gene–related molecular functions. To predict repurposable drugs for malaria treatment, the list of identified human disease candidate genes and their enriched biological processes generated from the analysis of human functional network were exploited in this chapter to investigate their contributions to other human–related diseases. This is to help predict repurposable treatment options that can be appropriated for malaria. Our analysis in this chapter was based on the hypothesis that varying combinations of disease–associated genes can influence the pathogenicity of similar diseases [281]. This hypothesis has been implemented in the analysis of gene expression datasets to identify breast cancer prognosis signatures and also to investigate disease similarities [281]. This is because, similarity between set of human diseases does not depend entirely on the shared disease-associated genes but rather, the biological processes influencing disease etiology. Having established earlier that proteins or genes within a biological network function through the modularity effect, it implies that they could contribute to specific processes within the system. Diseases are said to be similar if there are common biological processes that contribute to disease manifestation.

In this section, we implemented an implicit semantic similarity approach to investigate different diseases of the same disorder class as malaria. The purpose of this investigation is to measure the similarity between enriched gene ontology annotations of our identified host candidate genes (**Table 17**) and the annotations related to other diseases in order to predict treatment options or repurposable drugs that can be explored for malaria control. Similarity between two diseases is estimated by computationally quantifying the co-occurrence of associated ontological terms among the diseases other than exact gene matches [281].
The approach used in this chapter has been shown to perform better because it considers not only the exact biological processes that are common but rather a systematic approach to investigating semantic similarities between the gene ontology processes [97, 281].

6.1 Identifying Human Diseases in the Same Disorder Class with Malaria

Identifying similarities between diseases is dependent on the ability to explore genes or variants or biological processes shared among such diseases. In this section, we leveraged drug-target disease-associations to compute disease similarity.
Gene ontology annotations were retrieved from gene ontology database. We retrieved gene–disease associations from DisGeNET version 6.0 platform [282]. The platform contains about 628,685 gene-disease associations between 17,549 genes among approximately 24,166 diseases as at the time of our research.

Human disease ontology datasets were downloaded from disease ontology database [280]. The disease ontology terms are linked to terminologies that contain disease and disease-related concepts such as disease pathology (UMLS), MeSH, ICD-9 and ICD-10. The disease ontology dataset is to facilitate the cross-walk among disease-associated genes and the disease-related concepts. Pathway enrichment analysis of the human disease candidate genes (**Table 19**) elucidated diseases that the genes are involved. However, we investigated similarities between our human-malaria candidate genes and all human-disease genes in DisGeNet by filtering out all genes with no annotation and maintaining appropriate genes and their associated diseases.

To ascertain whether a disease is similar to malaria, we considered the semantic similarity score between the pair of diseases. The score is a quantitative measure of the underlying shared biological processes among the disease targets. A higher score between disease enriched processes suggests that the disease–pair and their associated candidate proteins are functionally similar thus, the likelihood for similar treatment options irrespective of the observed symptoms. The developed python–based model implemented for disease similarity uses our identified host targets (**Table 17**), disease–target datasets, gene ontology datasets as input data to predict similarity based on functional similarities inferred from associated gene ontology terms. With a defined threshold based on the upper quartile and interquartile range of the distribution given by tr = Q3+ϵ*IQR, where ϵ, tr, Q3 and IQR represent the tunning parameter ($0 \leq \epsilon \leq 1.5$), threshold, upper quartile and interquartile range respectively. The tuning parameter was set to 1.5 and the defined threshold was 0.47700375 for our analysis. **Figure 31** describes the semantic similarity scores between *Plasmodium falciparum* malaria and other diseases using enriched disease-associated processes.

Out of the 24,166 diseases in DisGeNet database, we identified 1,944 diseases to be semantically similar to malaria after defining a semantic similarity score threshold. We then filtered the disease hits by maintaining diseases whose targets are involved in the pathways of our host disease candidate key genes. As at the time of our studies, our host candidate genes were involved in 69 pathways (**Table 23**). We performed this analysis by estimating the Kappa and Jaccard statistics measure(described in equation 13 and equation 14 respectively) of a disease pair following by biological evidence from literature.

$$SimKPS(p,q) = \frac{\sigma_{pq} - \alpha_{pq}}{1 - \alpha_{pq}} \tag{13}$$

where σ_{pq} is defined as the observed frequency of co-occurrence between the profiles of protein p and q whereas α_{pq} is the likelihood of observing the profiles of protein p and q in the data under consideration [283].

$$SimUI(p,q) = \frac{|A_p \cap A_q|}{|A_p \cup A_q|} \tag{14}$$

where A_p and A_q represents the biological pathways associated to protein the profiles p and q.
The final filtered disease hits consisted of 115 as described in **Table 24**. These filtered disease hits identified to be similar to malaria are mostly pathogenic diseases including but not limited to parasitic, viral and bacterial infections that cause the human immune defense machinery to overproduce cytokines during host–pathogen interaction confirming the fact that malaria is an inflammatory disease [284]. These diseases fall in the category of mostly infectious, inflammatory and genetic neurological diseases which are caused by non–infectious agents. Among the top disease hits includes sickle cell anemia, liver dysfunction, fever, hepatitis and respiratory distress syndrome. The diseases described (**Table 24**) have been reported to be governed by same pathologic principles as malaria infection [284, 285].

Table 23. Predicted overall pathways associated with host candidate key proteins.

KEGG pathway ID	Pathway name	Class
path:hsa05140	Leishmaniasis	Infectious disease (parasitic)
path:hsa04064	NF-kappa B signaling pathway	Signal transduction
path:hsa01523	Antifolate resistance	
path:hsa04931	Insulin resistance	Endocrine and metabolic disease
path:hsa05169	Epstein-Barr virus infection	Infectious disease(viral)
path:hsa05152	Tuberculosis	Infectious disease(bacterial)
path:hsa05310	Asthma	Immune disease
path:hsa04933	AGE-RAGE signaling pathway in diabetic complications	Endocrine and metabolic disease
path:hsa04920	Adipocytokine signaling pathway	Endocrine system
path:hsa04350	TGF-beta signaling pathway	Signal transduction
path:hsa05150	Staphylococcus aureus infection	Infectious disease(bacterial)
path:hsa04668	TNF signaling pathway	Signal transduction
path:hsa05205	Proteoglycans in cancer	Cancer
path:hsa04940	Type I diabetes mellitus	Endocrine and metabolic disease
path:hsa04620	Toll-like receptor signaling pathway	Immune system
path:hsa05321	Inflammatory bowel disease (IBD)	Immune disease
path:hsa05130	Pathogenic Escherichia coli infection	Infectious disease(bacterial)
path:hsa05322	Systemic lupus erythematosus	Immune disease
path:hsa04630	JAK-STAT signaling pathway	Signal transduction
path:hsa05330	Allograft rejection	Immune disease
path:hsa05168	Herpes simplex virus 1 infection	Infectious disease(viral)
path:hsa05134	Legionellosis	Infectious disease(bacterial)
path:hsa04217	Necroptosis	Cell growth and death
path:hsa04514	Cell adhesion molecules	Signaling molecules and interaction
path:hsa05416	Viral myocarditis	Cardiovascular disease
path:hsa04621	NOD-like receptor signaling pathway	Immune system
path:hsa05133	Pertussis	Infectious disease(bacterial)

Continued on next page

Table 23 – continued from previous page

KEGG pathway ID	Pathway name	Class
path:hsa05144	Malaria	Infectious disease(parasitic)
path:hsa04071	Sphingolipid signaling pathway	Signal transduction
path:hsa05162	Measles	Infectious disease(viral)
path:hsa04151	PI3K-Akt signaling pathway	Signal transduction
path:hsa05146	Amoebiasis	Infectious disease(parasitic)
path:hsa04010	MAPK signaling pathway	Signal transduction
path:hsa05010	Alzheimer disease	Neurodegenerative disease
path:hsa04932	Non-alcoholic fatty liver disease (NAFLD)	Endocrine and metabolic disease
path:hsa04380	Osteoclast differentiation	Development and regeneration
path:hsa04060	Cytokine-cytokine receptor interaction	Signaling molecules and interaction
path:hsa05418	Fluid shear stress and atherosclerosis	Cardiovascular disease
path:hsa04657	IL-17 signaling pathway	Immune system
path:hsa05320	Autoimmune thyroid disease	Immune disease
path:hsa04150	mTOR signaling pathway	Signal transduction
path:hsa05143	African trypanosomiasis	Infectious disease(parasitic)
path:hsa05132	Salmonella infection	Infectious disease(bacterial)
path:hsa05165	Human papillomavirus infection	Infectious disease(viral)
path:hsa04930	Type II diabetes mellitus	Endocrine and metabolic disease
path:hsa04660	T cell receptor signaling pathway	Immune system
path:hsa04068	FoxO signaling pathway	Signal transduction
path:hsa04640	Hematopoietic cell lineage	Immune system
path:hsa04664	Fc epsilon RI signaling pathway	Immune system
path:hsa04622	RIG-I-like receptor signaling pathway	Immune system
path:hsa05414	Dilated cardiomyopathy (DCM)	Cardiovascular disease
path:hsa04670	Leukocyte transendothelial migration	Immune system
path:hsa04066	HIF-1 signaling pathway	Signal transduction
path:hsa05332	Graft-versus-host disease	Immune disease
path:hsa05142	Chagas disease	Infectious disease(parasitic)

Continued on next page

Table 23 – continued from previous page

KEGG pathway ID	Pathway name	Class
path:hsa05014	Amyotrophic lateral sclerosis (ALS)	Neurodegenerative disease
path:hsa05160	Hepatitis C	Infectious disease(viral)
path:hsa04612	Antigen processing and presentation	Immune system
path:hsa05410	Hypertrophic cardiomyopathy (HCM)	Cardiovascular disease
path:hsa04145	Phagosome	Transport and catabolism
path:hsa04650	Natural killer cell mediated cytotoxicity	Immune system
path:hsa04672	Intestinal immune network for IgA production	Immune system
path:hsa05145	Toxoplasmosis	Infectious disease
path:hsa05161	Hepatitis B	Infectious disease(viral)
path:hsa04210	Apoptosis	Cell growth and death
path:hsa05164	Influenza A	Infectious disease(viral)
path:hsa05323	Rheumatoid arthritis	Immune disease
path:hsa05166	Human T-cell leukemia virus 1 infection	Infectious disease(viral)
path:hsa05167	Kaposi sarcoma	Cancer

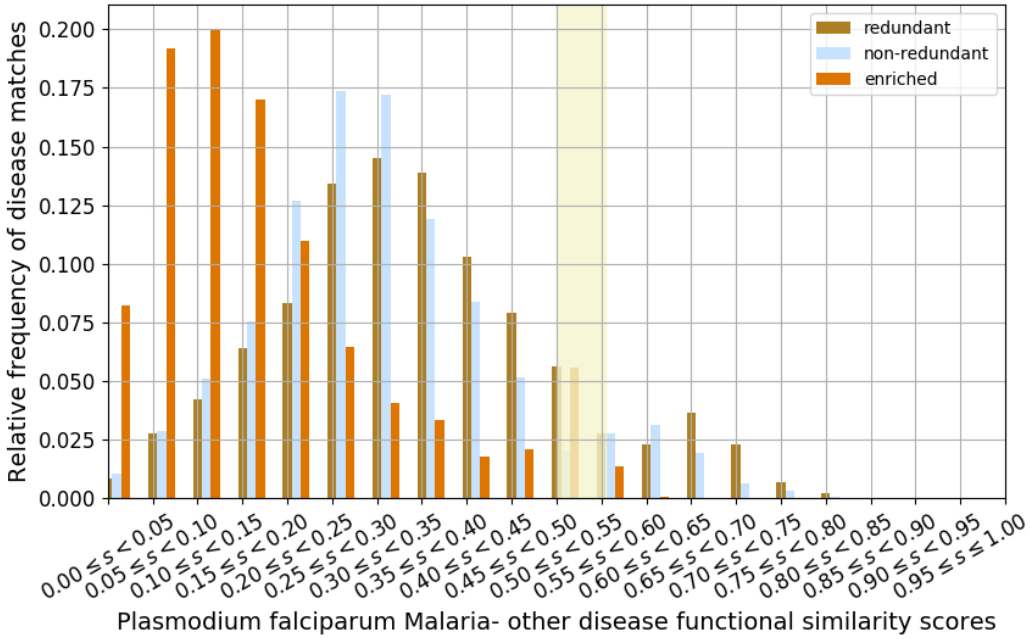

Figure 31. Different distributions of disease similarity scores obtained in terms of frequencies (proportions) of disease matches vs similarity scores between disease-associated processes. The bigger rectangular bar indicates the threshold for similarity between disease pair of which we used the enriched similarity score (ESS) for further analysis.

Table 24. Predicted malaria–similar diseases identified using semantic similarity approach. ESS represents the estimated enriched similarity scores.

Drug-ID	Disease Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0001175	Acquired Immunodeficiency Syndrome	226	69	0.54266	0.0	0.30531
C0002871	Anemia	277	69	0.51979	0.0	0.24909
C0002873	Anemia of chronic disease	140	66	0.51902	-0.04288	0.47143
C0002874	Aplastic Anemia	223	68	0.51064	-0.00899	0.30493
C0002893	Refractory anemias	248	69	0.49309	0.0	0.27823
C0002895	Anemia, Sickle Cell	255	68	0.55156	-0.00786	0.26667
C0003123	Anorexia	200	68	0.52716	-0.01002	0.34
C0003864	Arthritis	270	69	0.50061	0.0	0.25556
C0006111	Brain Diseases	250	69	0.52636	0.0	0.276
C0006118	Brain Neoplasms	277	69	0.51262	0.0	0.24909
C0006287	Bronchopulmonary Dysplasia	237	69	0.5246	0.0	0.29114
C0007785	Cerebral Infarction	263	69	0.50857	0.0	0.26236
C0007786	Brain Ischemia	237	69	0.52724	0.0	0.29114
C0007789	Cerebral Palsy	201	68	0.51322	-0.00997	0.33831
C0007847	Malignant tumor of cervix	283	69	0.51019	0.0	0.24382
C0011311	Dengue Fever	196	69	0.53769	0.0	0.35204

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0011991	Diarrhea	280	69	0.53301	0.0	0.24643
C0015672	Fatigue	261	69	0.52095	0.0	0.26437
C0015674	Chronic Fatigue Syndrome	148	69	0.56246	0.0	0.46622
C0015967	Fever	228	69	0.51222	0.0	0.30263
C0018621	Hay fever	178	69	0.55729	0.0	0.38764
C0019101	Hemorrhagic Fever with Renal Syndrome	109	68	0.54928	-0.01823	0.62385
C0019158	Hepatitis	268	69	0.50801	0.0	0.25746
C0019159	Hepatitis A	253	69	0.52661	0.0	0.27273
C0019187	Hepatitis, Alcoholic	178	68	0.55737	-0.01126	0.38202
C0019189	Hepatitis, Chronic	212	69	0.53011	0.0	0.32547
C0019193	Hepatitis, Toxic	253	69	0.52166	0.0	0.27273
C0019196	Hepatitis C	285	69	0.51194	0.0	0.24211
C0019207	Hepatoma, Morris	200	66	0.53196	-0.03021	0.33
C0019208	Hepatoma, Novikoff	200	66	0.53286	-0.03021	0.33
C0020542	Pulmonary Hypertension	214	69	0.52169	0.0	0.32243
C0021400	Influenza	269	69	0.50113	0.0	0.25651
C0023267	Fibroid Tumor	248	69	0.52817	0.0	0.27823
C0023290	Leishmaniasis, Visceral	182	69	0.54409	0.0	0.37912

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0023440	Acute Erythroblastic Leukemia	236	68	0.49545	-0.00849	0.28814
C0027497	Nausea	194	67	0.52302	-0.02071	0.34536
C0027498	Nausea and vomiting	219	67	0.52054	-0.01836	0.30594
C0027540	Necrosis	177	66	0.53155	-0.03411	0.37288
C0034063	Pulmonary Edema	186	67	0.53161	-0.02160	0.36022
C0034067	Pulmonary Emphysema	227	69	0.52845	0.0	0.30396
C0034069	Pulmonary Fibrosis	254	69	0.507	0.0	0.27165
C0035220	Respiratory Distress Syndrome, Newborn	193	67	0.53741	-0.02082	0.34715
C0035222	Respiratory Distress Syndrome, Adult	238	69	0.54066	0.0	0.28992
C0035235	Respiratory Syncytial Virus Infections	183	69	0.54611	0.0	0.37705
C0035242	Respiratory Tract Diseases	171	67	0.54095	-0.02348	0.39181
C0035436	Rheumatic Fever	166	68	0.51759	-0.01207	0.40964
C0036205	Sarcoidosis, Pulmonary	152	69	0.53268	0.0	0.45395
C0036974	Shock	106	66	0.5684	-0.05525	0.62264
C0036983	Septic Shock	113	68	0.5668	-0.01761	0.60177
C0038436	Post-Traumatic Stress Disorder	203	69	0.53531	0.0	0.33990
C0040034	Thrombocytopenia	269	69	0.51614	0.0	0.25651
C0041228	African Trypanosomiasis	134	67	0.57142	-0.02982	0.5

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0041296	Tuberculosis	277	69	0.50244	0.0	0.24909
C0041327	Tuberculosis, Pulmonary	207	69	0.54692	0.0	0.33333
C0041466	Typhoid Fever	125	67	0.56711	-0.03188	0.536
C0042721	Viral hepatitis	168	67	0.51131	-0.02389	0.39881
C0085293	Hepatitis E	121	68	0.55996	-0.01648	0.56198
C0085605	Liver Failure	239	69	0.5282	0.0	0.28870
C0085742	Injuries, Acute Brain	157	67	0.54269	-0.02555	0.42675
C0086404	Experimental Hepatoma	199	66	0.53361	-0.03036	0.33166
C0086565	Liver Dysfunction	250	69	0.49199	0.0	0.276
C0151332	Active tuberculosis	114	69	0.56211	0.0	0.60526
C0152171	Idiopathic pulmonary hypertension	202	67	0.48326	-0.01989	0.33168
C0155728	Other specified transient cerebral ischemias	212	68	0.5116	-0.00946	0.32075
C0206624	Hepatoblastoma	253	69	0.52569	0.0	0.27273
C0206754	Neuroendocrine Tumors	235	68	0.51095	-0.00853	0.28936
C0220620	Gastrointestinal Carcinoid Tumor	218	68	0.49072	-0.00919	0.31193
C0220650	Metastatic malignant neoplasm to brain	220	68	0.52063	-0.00911	0.30909
C0221505	Lesion of brain	191	69	0.53327	0.0	0.36126
C0231528	Myalgia	167	68	0.53991	-0.01199	0.40719

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0235946	Cerebral atrophy	268	69	0.48964	0.0	0.25746
C0241910	Hepatitis, Autoimmune	204	69	0.52297	0.0	0.33824
C0242584	Autoimmune thrombocytopenia	160	68	0.54667	-0.01252	0.425
C0242966	Systemic Inflammatory Response Syndrome	167	69	0.5554	0.0	0.41317
C0243026	Sepsis	280	69	0.5082	0.0	0.24643
C0270611	Brain Injuries	157	67	0.54269	-0.02555	0.42675
C0271650	Impaired glucose tolerance	273	69	0.52643	0.0	0.25275
C0271907	Acquired aplastic anemia	153	66	0.53914	-0.03935	0.43137
C0272945	Brain Lacerations	157	67	0.54269	-0.02555	0.42675
C0282687	Hemorrhagic Fever, Ebola	112	67	0.56975	-0.03533	0.59821
C0375023	Respiratory syncytial virus (RSV) infection in conditions classified elsewhere and of unspecified site	219	69	0.5294	0.0	0.31507
C0452047	Brain Injuries, Focal	157	67	0.54269	-0.02555	0.42675
C0521158	Recurrent tumor	258	67	0.53361	-0.01558	0.25969
C0524909	Hepatitis B, Chronic	230	69	0.51533	0.0	0.3
C0524910	Hepatitis C, Chronic	258	69	0.52971	0.0	0.26744
C0598935	Tumor Initiation	242	68	0.51708	-0.00829	0.28099

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C0600327	Toxic Shock Syndrome	135	67	0.53981	-0.02961	0.49629
C0740391	Middle Cerebral Artery Occlusion	204	68	0.51866	-0.00983	0.33333
C0740392	Infarction, Middle Cerebral Artery	209	68	0.54169	-0.00959	0.32536
C0751690	Malignant Peripheral Nerve Sheath Tumor	212	67	0.51253	-0.01896	0.31604
C0751955	Brain Infarction	166	67	0.52277	-0.02418	0.40361
C0917798	Cerebral Ischemia	252	69	0.50236	0.0	0.27381
C0917996	Cerebral Aneurysm	149	69	0.57101	0.0	0.46309
C0948008	Ischemic stroke	258	69	0.51378	0.0	0.26744
C1145670	Respiratory Failure	208	67	0.53727	-0.01932	0.32212
C1175175	Severe Acute Respiratory Syndrome	181	68	0.55008	-0.01107	0.37569
C1262760	Hepatitis, Drug-Induced	253	69	0.52172	0.0	0.27273
C1275126	TNF receptor-associated periodic fever syndrome (TRAPS)	141	67	0.55724	-0.02839	0.47528
C1282496	Metastasis from malignant tumor of prostate	228	68	0.51473	-0.00879	0.29825
C1290398	Cerebral arterial aneurysm	173	69	0.55418	0.0	0.39884
C1336708	Testicular Germ Cell Tumor	207	68	0.51821	-0.00969	0.32850
C1512409	Hepatocarcinogenesis	272	68	0.52216	-0.00737	0.25
C1519666	Tumor-Associated Vasculature	140	66	0.54477	-0.04288	0.47143

Continued on next page

Table 24 – continued from previous page

Disease-ID	Disease–Name	No. disease-target associated pathways	No. of common pathways	ESS	Kappa measure	Jaccard measure
C1519670	Tumor Angiogenesis	265	69	0.51521	0.0	0.26038
C1519680	Tumor Immunity	192	68	0.54724	-0.01044	0.35417
C1658953	tumor vasculature	196	68	0.52328	-0.01023	0.34694
C1719672	Severe Sepsis	194	69	0.53222	0.0	0.35567
C1800706	Idiopathic Pulmonary Fibrosis	245	69	0.54026	0.0	0.28163
C1857276	Trichohepatoenteric Syndrome	244	68	0.48569	-0.00822	0.27869
C3203102	Idiopathic pulmonary arterial hypertension	251	69	0.50683	0.0	0.27490
C3241937	Nonalcoholic Steatohepatitis	256	68	0.52614	-0.00783	0.26563
C3263723	Traumatic injury	104	67	0.54475	-0.03776	0.64423
C3469521	Fanconi anemia, Complementation group A (disorder)	246	69	0.52568	0.0	0.28049
C0264408	Childhood asthma	208	69	0.54101	0.0	0.33173
C0155877	Allergic asthma	211	69	0.5141	0.0	0.32701

6.2 Identifying Putative Drug Hits

We retrieved 1,426 approved drugs and their corresponding targets from DrugBank database out of which 1,282 were related to human. We first filtered these drugs by considering those with target–processes associated to malaria and predicted similar diseases (**Table 24**). We implemented the semantic approach described in (**Chapter 6 section 6.1**) to predict putative repurposable drugs. We analysed the existing drugs for treating the predicted similar diseases to elucidate those diseases whose drugs targets our predicted disease candidate genes.

From the identified drugs sharing some similarity in terms of processes, we extracted and ordered those that are over 1.5 of the interquartile range. This subset of drugs share higher similarity. With a defined similarity score threshold of 0.31099875 based on similarity in terms of processes the drugs are involved, we identified 26 drug hits (**Table 25**). Majority of the drug hits target interleukin, Interleukin-6, interferon (*IFN*)-γ and toll–like receptor as antagonist, agonist or inhibitors.
However, we observed that most of the drug hits involved in regulating host immune response to inflammatory–driven disorders targets the Tumor necrosis factor. Such drug hits are mostly used for the treatment of rheumatoid arthritis. Among the drug hits are chloroquine and certolizumab pegol.
Figure 32 shows the different distribution of relative frequency of drug matches against candidate gene–drug functional similarity scores.

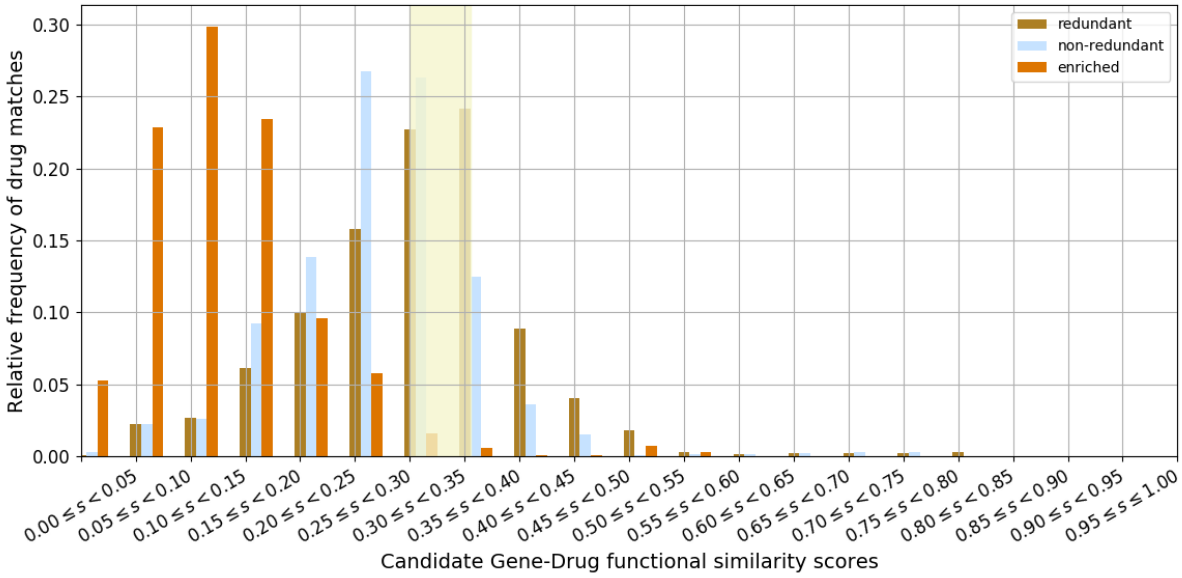

Figure 32. Distributions of drug similarity scores obtained in terms of relative frequency of drug matches against functional similarity scores between candidate gene and drug.

Table 25. Putative drug hits identified using semantic similarity approach.

Drug-ID	Drug-Name	ESS	NRSS	RSS
DB00608	Chloroquine	0.56954	0.7663	0.8063
DB08904	Certolizumab pegol	0.56393	0.7877	0.81003
DB06674	Golimumab	0.56393	0.7877	0.81003
DB00065	Infliximab	0.56393	0.7877	0.81003
DB00051	Adalimumab	0.53534	0.72775	0.75849
DB01296	Glucosamine	0.53079	0.69089	0.74441
DB00005	Etanercept	0.52932	0.72005	0.75654
DB08910	Pomalidomide	0.52833	0.69503	0.73025
DB00668	Epinephrine	0.52095	0.65483	0.69113
DB01041	Thalidomide	0.50493	0.64995	0.68472
DB01611	Hydroxychloroquine	0.40335	0.51466	0.58274
DB01250	Olsalazine	0.39014	0.47881	0.53864
DB05679	Ustekinumab	0.3856	0.4261	0.46755
DB06168	Canakinumab	0.37554	0.49231	0.55968
DB09036	Siltuximab	0.36098	0.48012	0.5382
DB01404	Ginseng	0.36071	0.4769	0.52545
DB01017	Minocycline	0.3551	0.46679	0.54531
DB08895	Tofacitinib	0.324	0.44454	0.50481
DB06273	Tocilizumab	0.31913	0.4223	0.47212
DB00029	Anistreplase	0.31537	0.40083	0.44003
DB00015	Reteplase	0.31537	0.40083	0.44003
DB00009	Alteplase	0.31537	0.40083	0.44003
DB00480	Lenalidomide	0.31461	0.41723	0.48212
DB00108	Natalizumab	0.31162	0.58395	0.62286
DB00031	Tenecteplase	0.31136	0.41706	0.4763
DB08877	Ruxolitinib	0.31119	0.44161	0.50625

6.3 Summary

The massive disease datasets present to the bioinformatics community have resulted in a continuous effort to understand diseases in an era whereby drug discovery is challenged by resistance and drug failure during clinical trials. Due to that, various computational methods and tools are implemented to harness the datasets to improve health care by translating biological knowledge into identifying drug targets and developing effective repurposable therapeutics thus, reducing both the drug development time and the risk of drug failure. In this chapter we implemented semantic similarity approach on the bases of exploring gene ontology annotations and processes associations between various human diseases to predict putative diseases in the same disorder class as malaria. The purpose of estimating disease similarity was to enable us predict diseases whose drugs can be appropriated experimentally for malaria treatment thus significantly reducing the cost and time involved in developing new drugs. We primarily used disease ontology, drug-bank datasets and disease–gene association datasets to carry out the analysis in this chapter. After defining the criteria of shared biological processes, pathways and literature-based evidence, we identified 115 similar diseases to malaria including but not limited to sickle cell anaemia, tuberculosis, respiratory distress, liver dysfunctioning and hepatitis as described in **Table 24**. We identified putative drug hits (**Table 25**) including but not limited to chloroquine, infliximab, Hydroxychloroquine, glucosamine, ginseng, minocycline, ruxolitinib and natalizumab which can be appropriated for malaria treatment. These drug hits have been reported to control malaria infection by inhibiting residual malaria infection, knocking parasite gene expression and activates eryptosis. Also, some of the drugs such as adalimumab, Natalizumab, etanercept, thalidomide, ustekinumab and canakinumab are anti-TNF monoclonal antibodies and anti-inflammatory agents that trigger immune response. Knowing that malaria is an inflammatory-response driven disease, the putative drug hits can undergo both computational and experimental repositioning for malaria treatment.

Chapter 7

7 General Discussion and Conclusion

Plasmodium falciparum malaria remains to be a major public health concern especially in highly endemic regions like Africa which contributes significantly to the global malaria morbidity and mortality statistics. The continuous spread of artemisinin resistance in Southeast Asia, a region known for founder events of chloroquine and sulphadoxine resistance, threatens the agenda of controlling and eradicating malaria [17]. Researchers have suggested that, the emergence of artemisinin parasite resistant strains in Africa would result in about 78 million additional cases annually [9]. Due to that, various research are conducted in the drug discovery field to develop effective new drugs.

In the beginning section of our study, we presented known antimalarial drugs, the origin of resistance and the resistant–associated variants rendering such drugs ineffective. The review of antimalarials provided an overview of their nature, mechanism of action and investigate those that are currently used for malaria treatment. We followed up by reviewing various computational methods implemented in drug discovery to complement experimental approaches. We emphasized on the role of machine learning, data mining, genomics and biological network analysis in current drug discovery pipeline. The purpose of the review was to investigate how to leverage computational approaches, disease data and drug datasets to predict targets, protein interactions within an interactome, enriched processes underlying disease pathology and repurposable drugs.

In this study, we aimed at elucidating potential possible mechanisms that may influence artemisinin reduced sensitivity or resistance in Africa. We conducted protein-based analysis to explore functional interactions between known parasite resistance genes and artemisinin drug targets to elucidate patterns that might contribute to drug resistance development. Interestingly, our results (**Figure 13**) showed that possible resistance to artemisinin may involve multiple parasite drug resistant genes such as *pfcrt, dhps, pfmdr1, dhfr* and *plasmepsin 2 (PF14_0077)*. This suggests the additive contribution of drug resistant genes and their functional relationship towards antimalarial resistance. Results provide gene/protein level functional interaction evidence on reported association of these drug resistant gene polymorphisms. These findings may support decisions on the nature of artemisinin resistance development involving multiple genes and the use of artemisinin combinations therapies particularly in hyper-endemic and hypo-endemic malaria regions. The results may suggest the likelihood of combining artemisinin with other agents or the use of other antimalarials. Such initiative may help increase drug sensitivity, thus, reducing the development and widespread of drug resistance.
Next, we performed further analysis to elucidate *pfk13* functional interactions that may contribute to possible host immune adaptiveness, artemisinin reduced sensitivity or malaria resistance in Africa by analyzing the assembled host-pathogen functional network. Network analysis (**Figure 25**) revealed that *pfk13* functionally interacts with other host regulatory proteins and kelch-like genes. This may suggest that the protein can influence and adapt to host immune response upon infection as well as drug resistance development under drug pressure.

Furthermore, we identified potential protein targets that can play essential role for developing effective antimalarial drugs and vaccines.
We conducted protein-based analysis by leveraging the various heterogeneous experimental and *in silico* datasets retrieved from databases and literature to assemble *Plasmodium falciparum*, human and human-*Plasmodium falciparum* functional protein-protein interaction network comprising of reviewed proteins. Using host–malaria GWAS summary statistics datasets from Kenya, Malawi and Gambia populations, we identified host-disease associated genes by mapping nominally significant SNPs to their associated genes. By leveraging Blonde et al. [113] clustering algorithm, we mapped these identified genes, malaria parasite selective variants and parasite variants under strong signature of selection unto the host and pathogen functional network respectively to partition the network into subnetworks (**Tables 11 and 16**). We observed that the subnetworks describe proteins involved in

similar processes.

Thereafter, we leveraged the topological features of nodes within the subnetworks of each assembled network to investigate nodes (candidate key proteins) that contribute significantly to the stability and integrity of the network. We identified hub genes and performed gene annotation and enrichment analysis to elucidate underlying statistically significant enriched biological processes and pathways of the genes that are involved. From the parasite assembled functional network, we predicted C6KTD2 *(SET1)* and C6KTB7 *(PFF1365c)* as essential target hubs mainly involved in the protein ubiquitination and histone-lysine methylation within the parasite. *SET1* is known to be involved in protein binding (GO:0005515, GO:0019904). C6KTB7 is mainly involved in ubiquitin-protein transferase activity (GO:0004842, GO:0019787) through the protein ubiquitination and modification pathway (UPA00143). Studies have shown that many biological processes and substrates are targeted by the ubiquitin pathway such that instability or modification in ubiquitination and deubiquitination reactions influences the pathogenesis of many eukaryotic system related diseases. For instance, the dysregulation of ubiquitin ligase is associated to neurodegenerative disorders such as Parkinson's disease and infectious diseases including tuberculosis. These targets have been reported as candidates for drug and vaccine development. Our results confirm the essentiality of these targets. Also, our analysis showed that these targets could be critical for combinatorial drug design.

Interactome analysis on the host functional network revealed P22301 *(IL10)*, P05362 *(ICAM1)*, P01375 *(TNF)*, P30480 *(HLA-B)*, P16284 *(PECAM1)* and O00206 *(TLR4)* as key targets. Interestingly, we observed that HLA-B encoded multiple proteins within the host network. This suggests that HLA-B has multiple protein coding exons. This observation may suggest the viability of the gene in Africa and its contribution towards malaria susceptibility. However, we acknowledge the role of these hub genes in other populations, but we have insufficient data to explicitly say that there is a difference. These host candidate key proteins are involved in immune response and resistance against malaria infection including severe and cerebral malaria, thus, critical targets for adjunctive and antibody-based therapy for malaria control [286, 287, 288]. Studies have shown that even the most potent artemisinin derivatives in insufficient to treat severe malaria and cerebral malaria and therefore requires the additive contribution of host-directed therapy involved in modulating host response to infection. This may contribute significantly to improve treatment efficacy, reduce disease-associated complexity, reduce malaria-associated mortality and morbidity as well as slow artemisinin resistance development. Functional analysis revealed 23 significantly enriched malaria-related biological processes described in (**Table 18**) were identified. Enrichment analysis showed an overlap enriched pathways to other infectious diseases, including tuberculosis, measles, leishmaniasis, and hepatitis (**Table 23**). These findings support the evidence of similarity between diseases and overlapping gene processes underlying such disorders.

Next, we investigated the shortest paths to elucidate pathways that could account for parasite adaptiveness to host response and drug resistance development (**Tables 21 and 22**). We investigated the pathways of immune tolerance and potential resistance development among the host and pathogen key targets by analyzing shortest distance between these genes with the host–*Plasmodium falciparum* functional network. Our analysis showed that these shortest paths between the candidate genes or proteins are mediated by host genes involved in cell regulatory activities, inflammatory response and general cell integrity.(**Figures 29** and **30**).

Additionally, we investigated the functional interactions between reported artemisinin drug targets and *Plasmodium falciparum* resistance selective genes to explore functional interactions and mechanisms that can contribute to artemisinin resistance in Africa. We observed that, these selective variants functionally interact with each other to form hubs whereas artemisinin targets functionally associates to these hubs either by high degree or betweenness **(Figure13)**. Also, we explored to understand modes of resistance development and realized that, this could arise within the African populations due to interference by pathogen candidate targets inhibiting host genes conferring immunity to malaria.

Finally, we implemented semantic similarity approach to identify 115 diseases similar to malaria

(**Table 24**) that facilitate the prediction of 26 repurposable drug hits (**Table 25**) that can be computationally and experimentally modified for malaria treatment.

Critically, as for any other computational or in silico approaches which always need validation through further functional study, we believe that the presented in silico (computational approach and pipeline) can inform functional study for potential experimental and clinical validation.

7.1 Potential Impact

Identifying novel potential drugs and repurposable drugs could contribute to developing informed hypothesis for malaria drug research. Such hypothesis could provide insights for drug repurposing and/or discovery of highly efficient therapeutics and vaccines, thus, streamlining experimental approaches and reducing significantly the cost and drug production time.

In this study, protein-based analysis between pathogen, host and host-pathogen has contributed to investigating functional interactions underlying disease pathogenesis and host response to infection. Such analysis is critical to understand the genetic architecture of complex biological phenomena and the enriched pathways and biological processes relevant for prioritizing experimental findings.

Analysis on potential resistance to current malaria drug, artemisinin, has provided insights that could contribute to developing informed public health policies that would on the use of artemisinin combination therapies particularly in endemic settings. This may contribute to the longevity of the drug.

Malaria is reported to have significant measurable direct and indirect costs, thus, shown to be a major constraint to economic development by retarding productivity and growth. This is experienced mainly through diversion of resources to control malaria and loss of human lives which is very essential to the socio-economic development of a nation. This study has therefore contributed to knowledge that could help in the translation of research findings in measurable outputs for malaria control.

7.2 Potential Implementation Strategies

Perturbation analysis is critical to investigate targets that exacerbate or hinder a disease. In-vitro gene knock-out analysis on the identified targets is recommended because this could help determine their mechanistic function, expression level and how it can be engineered and/or harnessed in drug research. We further suggest in vivo studies to investigate the targets druggability.
We propose experimental validation of predicted drug resistant mechanisms to fully ascertain the functional effect of protein-protein interactions and their systemic effect.

7.3 Limitations and Future Work

In this project, we based our functional networks on generated protein–protein interaction datasets to perform our analysis between human (host) and parasite (*Plasmodium falciparum*). We have used the protein–protein interaction datasets to assemble *Plasmodium falciparum*, *Plasmodium falciparum* malaria–specific, human and human–parasite functional networks in order to achieve results. Nonetheless, a way to improve upon this studies is to explicitly include micro-array datasets and other complementary biological datasets such as transcriptomic datasets and time series gene expression data in order to comprehensively understand gene expression changes, thus improving our knowledge on human-*falciparum* interactions to fully perform advanced computational analysis. This will provide an insightful understanding of the conditions at which the observed functional interactions would occur.
As at the time of our study, we observed that majority of *Plasmodium falciparum* protein data from uniprot are uncharacterized. A way to improve on our study could be the application of semantic

similarity-based approaches to functionally annotate uncharacterized parasite proteins. Such annotation would contribute to increased functional datasets thus increasing power and coverage for analysis. This approach could improve upon our analysis on exploring interactions between parasite selective variants and existing drug targets to elucidate their role to resistance development.
Overall, our current study results have provided insights to the dynamic patterns of human-*Plasmodium falciparum* functional network and the mechanisms involved in disease pathogenesis and possible resistance development paths.

Appendix

Table A1. Host Malaria–specific genes and other host genes predicted from functional interaction generated from genemania database.

Gene	Description
OR51V1	olfactory receptor family 51 subfamily V member 1
NANOS2	nanos C2HC-type zinc finger 2
FREM3	FRAS1 related extracellular matrix 3
BGLAP	bone gamma-carboxyglutamate protein
MARVELD3	MARVEL domain containing 3
SCO1	SCO1 cytochrome c oxidase assembly protein
GYPB	glycophorin B
IL4	interleukin 4
USP38	ubiquitin specific peptidase 38
GYPA	glycophorin A (MNS blood group)
ATP2B4	ATPase plasma membrane Ca2+ transporting 4
INPP4B	inositol polyphosphate-4-phosphatase type II B
HBE1	hemoglobin subunit epsilon 1
HLA-DRB5	major histocompatibility complex, class II, DR beta 5
G6PD	glucose-6-phosphate dehydrogenase
AMH	anti-Mullerian hormone
TLR9	toll like receptor 9
IL13	interleukin 13
TLR4	toll like receptor 4
MYH3	myosin, heavy chain 3, skeletal muscle, embryonic
HP	haptoglobin
CR1	complement component 3b/4b receptor 1 (Knops blood group)
HLA-B	major histocompatibility complex, class I, B
FCGR3B	Fc fragment of IgG receptor IIIb
FCGR2A	Fc fragment of IgG receptor IIa
IL10	interleukin 10

Continued on next page

Table A1 – continued from previous page

Gene	Description
HBB	hemoglobin subunit beta
IL1RN	interleukin 1 receptor antagonist
FCGR3A	Fc fragment of IgG receptor IIIa
MBL2	mannose binding lectin 2
TNFAIP1	TNF alpha induced protein 1
PSMB9	proteasome subunit beta 9
DDC	dopa decarboxylase
ACKR1	atypical chemokine receptor 1
MIF	macrophage migration inhibitory factor (glycosylation-inhibiting factor)
HMOX1	heme oxygenase 1
CD36	CD36 molecule
TNF	tumor necrosis factor
NOS2	nitric oxide synthase 2
IL1B	interleukin 1 beta
IRF1	interferon regulatory factor 1
ICAM1	intercellular adhesion molecule 1
HMOX2	heme oxygenase 2
NANOS3	nanos C2HC-type zinc finger 3
NANOS1	nanos C2HC-type zinc finger 1
DDTL	D-dopachrome tautomerase-like
SCARB1	scavenger receptor class B member 1
HMX3	H6 family homeobox 3
DDT	D-dopachrome tautomerase
CXCL1	C-X-C motif chemokine ligand 1
HBD	hemoglobin subunit delta
HBG1	hemoglobin subunit gamma 1
FREM2	FRAS1 related extracellular matrix protein 2
APCS	amyloid P component, serum
IL10RA	interleukin 10 receptor subunit alpha
IL1R2	interleukin 1 receptor type 2

Continued on next page

Table A1 – continued from previous page

Gene	Description
HBG2	hemoglobin subunit gamma 2
SELT	selenoprotein T
CR2	complement component 3d receptor 2
SCARB2	scavenger receptor class B member 2
GYPC	glycophorin C (Gerbich blood group)
MGP	matrix Gla protein

References

1. Veiga MI, Ferreira PE, Jörnhagen L, Malmberg M, Kone A, et al. (2011) Novel polymorphisms in plasmodium falciparum abc transporter genes are associated with major act antimalarial drug resistance. PloS one 6: e20212.

2. Driss A, Hibbert JM, Wilson NO, Iqbal SA, Adamkiewicz TV, et al. (2011) Genetic polymorphisms linked to susceptibility to malaria. Malaria journal 10: 271.

3. Organization WH, et al. (2015) Control and elimination of Plasmodium vivax malaria: a technical brief. World Health Organization.

4. Organization WH (2016) World malaria report 2015. World Health Organization.

5. Organization WH, et al. (2017) World malaria report 2017 .

6. De Kock M, Tarning J, Workman L, Nyunt M, Adam I, et al. (2017) Pharmacokinetics of sulfadoxine and pyrimethamine for intermittent preventive treatment of malaria during pregnancy and after delivery. CPT: pharmacometrics & systems pharmacology 6: 430–438.

7. Barber BE, Rajahram GS, Grigg MJ, William T, Anstey NM (2017) World malaria report: time to acknowledge plasmodium knowlesi malaria. Malaria journal 16: 135.

8. Le Bras J, Durand R (2003) The mechanisms of resistance to antimalarial drugs in plasmodium falciparum. Fundamental & clinical pharmacology 17: 147–153.

9. Ouji M, Augereau JM, Paloque L, Benoit-Vical F (2018) Plasmodium falciparum resistance to artemisinin-based combination therapies: A sword of damocles in the path toward malaria elimination. Parasite 25.

10. Molina-Cruz A, Barillas-Mury C (2014) The remarkable journey of adaptation of the plasmodium falciparum malaria parasite to new world anopheline mosquitoes. Memórias do Instituto Oswaldo Cruz 109: 662–667.

11. Kats LM, Fernandez KM, Glenister FK, Herrmann S, Buckingham DW, et al. (2014) An exported kinase (fikk4. 2) that mediates virulence-associated changes in plasmodium falciparum-infected red blood cells. International journal for parasitology 44: 319–328.

12. Wilson ME, Kantele A, Jokiranta TS (2011) Review of cases with the emerging fifth human malaria parasite, plasmodium knowlesi. Clinical infectious diseases 52: 1356–1362.

13. Hayton K, Su Xz (2008) Drug resistance and genetic mapping in plasmodium falciparum. Current genetics 54: 223–239.

14. Ramaprasad A, Pain A, Ravasi T (2012) Defining the protein interaction network of human malaria parasite plasmodium falciparum. Genomics 99: 69–75.

15. Chen Q, Schlichtherle M, Wahlgren M (2000) Molecular aspects of severe malaria. Clinical microbiology reviews 13: 439–450.

16. Okombo J, Chibale K (2018) Recent updates in the discovery and development of novel antimalarial drug candidates. MedChemComm 9: 437–453.

17. Antony HA, Parija SC (2016) Antimalarial drug resistance: An overview. Tropical parasitology 6: 30.

18. malERA Refresh Consultative Panel on Insecticide, Resistance D (2017) malera: An updated research agenda for insecticide and drug resistance in malaria elimination and eradication. PLoS medicine 14: e1002450.

19. Dieye B, Affara M, Sangare L, Joof F, Ndiaye YD, et al. (2016) West africa international centers of excellence for malaria research: Drug resistance patterns to artemether–lumefantrine in senegal, mali, and the gambia. The American journal of tropical medicine and hygiene 95: 1054–1060.

20. Raj DK, Mu J, Jiang H, Kabat J, Singh S, et al. (2009) Disruption of a plasmodium falciparum multidrug resistance-associated protein (pfmrp) alters its fitness and transport of antimalarial drugs and glutathione. Journal of Biological Chemistry 284: 7687–7696.

21. Kim Y, Schneider K (2013) Evolution of drug resistance in malaria parasite populations. Nature Education Knowledge 4: 6.

22. Imwong M, Suwannasin K, Kunasol C, Sutawong K, Mayxay M, et al. (2017) The spread of artemisinin-resistant plasmodium falciparum in the greater mekong subregion: a molecular epidemiology observational study. The Lancet Infectious Diseases 17: 491–497.

23. White NJ (2004) Antimalarial drug resistance. The Journal of clinical investigation 113: 1084–1092.

24. Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, et al. (2008) Evidence of artemisinin-resistant malaria in western cambodia. New England Journal of Medicine 359: 2619–2620.

25. Rono MK, Nyonda MA, Simam JJ, Ngoi JM, Mok S, et al. (2018) Adaptation of plasmodium falciparum to its transmission environment. Nature ecology & evolution 2: 377.

26. Rapanoel HA, Mazandu GK, Mulder NJ (2013) Predicting and analyzing interactions between mycobacterium tuberculosis and its human host. PLoS One 8: e67472.

27. Vaughan AM, Aly AS, Kappe SH (2008) Malaria parasite pre-erythrocytic stage infection: gliding and hiding. Cell host & microbe 4: 209–218.

28. Aly AS, Vaughan AM, Kappe SH (2009) Malaria parasite development in the mosquito and infection of the mammalian host. Annual review of microbiology 63: 195–221.

29. Mohandas N, An X (2012) Malaria and human red blood cells. Medical microbiology and immunology 201: 593–598.

30. Djimdé A, Doumbo OK, Cortese JF, Kayentao K, Doumbo S, et al. (2001) A molecular marker for chloroquine-resistant falciparum malaria. New England journal of medicine 344: 257–263.

31. Piedade R, Gil JP (2011) The pharmacogenetics of antimalaria artemisinin combination therapy. Expert opinion on drug metabolism & toxicology 7: 1185–1200.

32. Croft SL, Duparc S, Arbe-Barnes SJ, Craft JC, Shin CS, et al. (2012) Review of pyronaridine anti-malarial properties and product characteristics. Malaria journal 11: 270.

33. Ringwald P, Bickii J, Basco LK (1998) Efficacy of oral pyronaridine for the treatment of acute uncomplicated falciparum malaria in african children. Clinical infectious diseases 26: 946–953.

34. Benjamin J, Moore B, Lee ST, Senn M, Griffin S, et al. (2012) Artemisinin-naphthoquine combination therapy for uncomplicated pediatric malaria: A tolerability, safety and preliminary efficacy study. Antimicrobial agents and chemotherapy : AAC–06248.

35. Basso LG, Rodrigues RZ, Naal RM, Costa-Filho AJ (2011) Effects of the antimalarial drug primaquine on the dynamic structure of lipid model membranes. Biochimica et Biophysica Acta (BBA)-Biomembranes 1808: 55–64.

36. Baird JK, Hoffman SL (2004) Primaquine therapy for malaria. Clinical infectious diseases 39: 1336–1345.

37. Masters BR (2016). Mandell, douglas, and bennett's principles and practice of infectious diseases, (2015) eds: John e. bennett, raphael dolin, martin j. blaser. isbn: 13-978-1-4557-4801-3, elsevier saunders.

38. White N, Warrell D, Bunnag D, Looareesuwan S, Chongsuphajaisiddhi T, et al. (1981) Quinidine in falciparum malaria. The Lancet 318: 1069–1071.

39. Ramakrishnan G, Chandra N, Srinivasan N (2017) Exploring anti-malarial potential of fda approved drugs: an in silico approach. Malaria journal 16: 290.

40. Ginsburg H, Abdel-Haleem AM (2016) Malaria parasite metabolic pathways (mpmp) upgraded with targeted chemical compounds. Trends in parasitology 32: 7–9.

41. McGready R, Cho T, Villegas L, Brockman A, van Vugt M, et al. (2001) Randomized comparison of quinine-clindamycin versus artesunate in the treatment of falciparum malaria in pregnancy. Transactions of the Royal Society of Tropical Medicine and Hygiene 95: 651–656.

42. Stejskal F, Nohỳnková E, Kosina P, Kulichová J (2018) Diagnosis, treatment and prophylaxis of malaria in the czech republic. Klinicka mikrobiologie a infekcni lekarstvi 24: 20–30.

43. Srivastava IK, Vaidya AB (1999) A mechanism for the synergistic antimalarial action of atovaquone and proguanil. Antimicrobial agents and chemotherapy 43: 1334–1339.

44. Fidock DA, Nomura T, Talley AK, Cooper RA, Dzekunov SM, et al. (2000) Mutations in the p. falciparum digestive vacuole transmembrane protein pfcrt and evidence for their role in chloroquine resistance. Molecular cell 6: 861–871.

45. Hayward R, Saliba KJ, Kirk K (2005) pfmdr1 mutations associated with chloroquine resistance incur a fitness cost in plasmodium falciparum. Molecular microbiology 55: 1285–1295.

46. Roper C, Pearce R, Nair S, Sharp B, Nosten F, et al. (2004) Intercontinental spread of pyrimethamine-resistant malaria. Science 305: 1124–1124.

47. Sridaran S, McClintock SK, Syphard LM, Herman KM, Barnwell JW, et al. (2010) Anti-folate drug resistance in africa: meta-analysis of reported dihydrofolate reductase (dhfr) and dihydropteroate synthase (dhps) mutant genotype frequencies in african plasmodium falciparum parasite populations. Malaria journal 9: 247.

48. Witkowski B, Duru V, Khim N, Ross LS, Saintpierre B, et al. (2017) A surrogate marker of piperaquine-resistant plasmodium falciparum malaria: a phenotype–genotype association study. The Lancet Infectious Diseases 17: 174–183.

49. Gupta B, Xu S, Wang Z, Sun L, Miao J, et al. (2014) Plasmodium falciparum multidrug resistance protein 1 (pfmrp1) gene and its association with in vitro drug susceptibility of parasite isolates from north-east myanmar. Journal of Antimicrobial Chemotherapy 69: 2110–2117.

50. Duru V, Khim N, Leang R, Kim S, Domergue A, et al. (2015) Plasmodium falciparum dihydroartemisinin-piperaquine failures in cambodia are associated with mutant k13 parasites presenting high survival rates in novel piperaquine in vitro assays: retrospective and prospective investigations. BMC medicine 13: 305.

51. Agrawal S, Moser KA, Morton L, Cummings MP, Parihar A, et al. (2017) Association of a novel mutation in the plasmodium falciparum chloroquine resistance transporter with decreased piperaquine sensitivity. The Journal of infectious diseases 216: 468–476.

52. Mungthin M, Watanatanasup E, Sitthichot N, Suwandittakul N, Khositnithikul R, et al. (2017) Influence of the pfmdr1 gene on in vitro sensitivities of piperaquine in thai isolates of plasmodium falciparum. The American journal of tropical medicine and hygiene 96: 624–629.

53. Labadie-Bracho M, Adhin MR (2013) Increased pfmdr1 copy number in plasmodium falciparum isolates from suriname. Tropical Medicine & International Health 18: 796–799.

54. Nagasundaram N, Chakraborty C, Karthick V, Balaji V, Siva R, et al. (2016) Mechanism of artemisinin resistance for malaria pfatp6 l263 mutations and discovering potential antimalarials: An integrated computational approach. Scientific reports 6: 30106.

55. Kerb R, Fux R, Mörike K, Kremsner PG, Gil JP, et al. (2009) Pharmacogenetics of antimalarial drugs: effect on metabolism and transport. The Lancet infectious diseases 9: 760–774.

56. Mehlotra RK, Fujioka H, Roepe PD, Janneh O, Ursos LM, et al. (2001) Evolution of a unique plasmodium falciparum chloroquine-resistance phenotype in association with pfcrt polymorphism in papua new guinea and south america. Proceedings of the National Academy of Sciences 98: 12689–12694.

57. Takala-Harrison S, Laufer MK (2015) Antimalarial drug resistance in africa: key lessons for the future. Annals of the New York Academy of Sciences 1342: 62–67.

58. PANDEY AV, Bisht H, Babbarwal VK, Srivastava J, Pandey KC, et al. (2001) Mechanism of malarial haem detoxification inhibition by chloroquine. Biochemical Journal 355: 333–338.

59. Sidhu ABS, Verdier-Pinard D, Fidock DA (2002) Chloroquine resistance in plasmodium falciparum malaria parasites conferred by pfcrt mutations. Science 298: 210–213.

60. Tewari SG, Prigge ST, Reifman J, Wallqvist A (2017) Using a genome-scale metabolic network model to elucidate the mechanism of chloroquine action in plasmodium falciparum. International Journal for Parasitology: Drugs and Drug Resistance 7: 138–146.

61. Cooper RA, Lane KD, Deng B, Mu J, Patel JJ, et al. (2007) Mutations in transmembrane domains 1, 4 and 9 of the plasmodium falciparum chloroquine resistance transporter alter susceptibility to chloroquine, quinine and quinidine. Molecular microbiology 63: 270–282.

62. Gatton ML, Martin LB, Cheng Q (2004) Evolution of resistance to sulfadoxine-pyrimethamine in plasmodium falciparum. Antimicrobial agents and chemotherapy 48: 2116–2123.

63. Abdul-Ghani R, Farag HF, Allam AF (2013) Sulfadoxine-pyrimethamine resistance in plasmodium falciparum: a zoomed image at the molecular level within a geographic context. Acta tropica 125: 163–190.

64. McCollum AM, Basco LK, Tahar R, Udhayakumar V, Escalante AA (2008) Hitchhiking and selective sweeps of plasmodium falciparum sulfadoxine and pyrimethamine resistance alleles in a population from central africa. Antimicrobial agents and chemotherapy 52: 4089–4097.

65. Mbugi EV, Mutayoba BM, Malisa AL, Balthazary ST, Nyambo TB, et al. (2006) Drug resistance to sulphadoxine-pyrimethamine in plasmodium falciparum malaria in mlimba, tanzania. Malaria journal 5: 94.

66. Vinayak S, Alam MT, Mixson-Hayden T, McCollum AM, Sem R, et al. (2010) Origin and evolution of sulfadoxine resistant plasmodium falciparum. PLoS pathogens 6: e1000830.

67. McCollum AM, Schneider KA, Griffing SM, Zhou Z, Kariuki S, et al. (2012) Differences in selective pressure on dhps and dhfr drug resistant mutations in western kenya. Malaria journal 11: 77.

68. McCollum AM, Poe AC, Hamel M, Huber C, Zhou Z, et al. (2006) Antifolate resistance in plasmodium falciparum: multiple origins and identification of novel dhfr alleles. The Journal of infectious diseases 194: 189–197.

69. Roper C, Pearce R, Bredenkamp B, Gumede J, Drakeley C, et al. (2003) Antifolate antimalarial resistance in southeast africa: a population-based analysis. The Lancet 361: 1174–1181.

70. Hoglund RM, Workman L, Edstein MD, Thanh NX, Quang NN, et al. (2017) Population pharmacokinetic properties of piperaquine in falciparum malaria: an individual participant data meta-analysis. PLoS medicine 14: e1002212.

71. Mukherjee A, Gagnon D, Wirth DF, Richard D (2018) Inactivation of plasmepsins 2 and 3 sensitizes plasmodium falciparum to the antimalarial drug piperaquine. Antimicrobial agents and chemotherapy 62: e02309–17.

72. Price RN, Uhlemann AC, Brockman A, McGready R, Ashley E, et al. (2004) Mefloquine resistance in plasmodium falciparum and increased pfmdr1 gene copy number. The Lancet 364: 438–447.

73. Preechapornkul P, Imwong M, Chotivanich K, Pongtavornpinyo W, Dondorp AM, et al. (2009) Plasmodium falciparum pfmdr1 amplification, mefloquine resistance, and parasite fitness. Antimicrobial agents and chemotherapy 53: 1509–1515.

74. Suresh N, Haldar K (2018) Mechanisms of artemisinin resistance in plasmodium falciparum malaria. Current opinion in pharmacology 42: 46–54.

75. Sirima SB, Ogutu B, Lusingu JP, Mtoro A, Mrango Z, et al. (2016) Comparison of artesunate–mefloquine and artemether–lumefantrine fixed-dose combinations for treatment of uncomplicated plasmodium falciparum malaria in children younger than 5 years in sub-saharan africa: a randomised, multicentre, phase 4 trial. The Lancet Infectious Diseases 16: 1123–1133.

76. Nawaz F, Nsobya SL, Kiggundu M, Joloba M, Rosenthal PJ (2009) Selection of parasites with diminished drug susceptibility by amodiaquine-containing antimalarial regimens in uganda. The Journal of infectious diseases 200: 1650–1657.

77. Takala-Harrison S, Jacob CG, Arze C, Cummings MP, Silva JC, et al. (2014) Independent emergence of artemisinin resistance mutations among plasmodium falciparum in southeast asia. The Journal of infectious diseases 211: 670–679.

78. Kamau E, Campino S, Amenga-Etego L, Drury E, Ishengoma D, et al. (2014) K13-propeller polymorphisms in plasmodium falciparum parasites from sub-saharan africa. The Journal of infectious diseases 211: 1352–1355.

79. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, et al. (2014) A molecular marker of artemisinin-resistant plasmodium falciparum malaria. Nature 505: 50.

80. Baraka V, Mavoko HM, Nabasumba C, Francis F, Lutumba P, et al. (2018) Impact of treatment and re-treatment with artemether-lumefantrine and artesunate-amodiaquine on selection of plasmodium falciparum multidrug resistance gene-1 polymorphisms in the democratic republic of congo and uganda. PloS one 13: e0191922.

81. Somé AF, Séré YY, Dokomajilar C, Zongo I, Rouamba N, et al. (2010) Selection of known plasmodium falciparum resistance-mediating polymorphisms by artemether-lumefantrine and amodiaquine-sulfadoxine-pyrimethamine but not dihydroartemisinin-piperaquine in burkina faso. Antimicrobial agents and chemotherapy 54: 1949–1954.

82. Ataide R, Ashley EA, Powell R, Chan JA, Malloy MJ, et al. (2017) Host immunity to plasmodium falciparum and the assessment of emerging artemisinin resistance in a multinational cohort. Proceedings of the National Academy of Sciences 114: 3515–3520.

83. Network MGE (2019) Insights into malaria susceptibility using genome-wide data on 17,000 individuals from africa, asia and oceania. Nature communications 10.

84. Amor A, Toro C, Fernández-Martínez A, Baquero M, Benito A, et al. (2012) Molecular markers in plasmodium falciparum linked to resistance to anti-malarial drugs in samples imported from africa over an eight-year period (2002-2010): impact of the introduction of artemisinin combination therapy. Malaria journal 11: 100.

85. Miraclin TA, Matthew A, Rupali P (2016) Decreased response to artemisinin combination therapy in falciparum malaria: A preliminary report from south india. Tropical parasitology 6: 85.

86. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, et al. (2009) Genome-wide and fine-resolution association analysis of malaria in west africa. Nature genetics 41: 657.

87. Timmann C, Thye T, Vens M, Evans J, May J, et al. (2012) Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 489: 443.

88. Network MGE (2015) A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 526: 253.

89. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, et al. (2013) Imputation-based meta-analysis of severe malaria in three african populations. PLoS genetics 9: e1003509.

90. Network TMGE (2008) A global network for investigating the genomic epidemiology of malaria. Nature 456: 732.

91. Consortium GP, et al. (2015) A global reference for human genetic variation. Nature 526: 68.

92. Chimusa ER, Mbiyavanga M, Mazandu GK, Mulder NJ (2015) ancgwas: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. Bioinformatics 32: 549–556.

93. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. (2004) Uniprot: the universal protein knowledgebase. Nucleic acids research 32: D115–D119.

94. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, et al. (2005) A protein interaction network of the malaria parasite plasmodium falciparum. Nature 438: 103.

95. Bossi A, Lehner B (2009) Tissue specificity and the human protein interaction network. Molecular systems biology 5: 260.

96. Croft D, O'kelly G, Wu G, Haw R, Gillespie M, et al. (2010) Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 39: D691–D697.

97. Mazandu GK, Chimusa ER, Rutherford K, Zekeng EG, Gebremariam ZZ, et al. (2017) Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. Briefings in bioinformatics 19: 1141–1152.

98. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2008) Interpro: the integrative protein signature database. Nucleic acids research 37: D211–D215.

99. Consortium GO (2004) The gene ontology (go) database and informatics resource. Nucleic acids research 32: D258–D261.

100. Aoki KF, Kanehisa M (2005) Using the kegg database resource. Current protocols in bioinformatics 11: 1–12.

101. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2007) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic acids research 36: D623–D631.

102. Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) String: a database of predicted functional associations between proteins. Nucleic acids research 31: 258–261.

103. Wuchty S (2006) Topology and weights in a protein domain interaction network–a novel way to predict protein interactions. BMC genomics 7: 122.

104. Wuchty S, Ipsaro JJ (2007) A draft of protein interactions in the malaria parasite p. falciparum. Journal of proteome research 6: 1461–1470.

105. Wuchty S, Adams JH, Ferdig MT (2009) A comprehensive plasmodium falciparum protein interaction map reveals a distinct architecture of a core interactome. Proteomics 9: 1841–1849.

106. Wuchty S (2007) Rich-club phenomenon in the interactome of p. falciparum—artifact or signature of a parasitic life style? PloS one 2: e335.

107. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, et al. (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301: 1503–1508.

108. Young JA, Fivelman QL, Blair PL, de la Vega P, Le Roch KG, et al. (2005) The plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification. Molecular and biochemical parasitology 143: 67–79.

109. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2011) The intact molecular interaction database in 2012. Nucleic acids research 40: D841–D846.

110. Chatr-Aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2006) Mint: the molecular interaction database. Nucleic acids research 35: D572–D574.

111. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, et al. (2017) The biogrid interaction database: 2017 update. Nucleic acids research 45: D369–D379.

112. Agüero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, et al. (2008) Genomic-scale prioritization of drug targets: the tdr targets database. Nature reviews Drug discovery 7: 900.

113. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008: P10008.

114. Zhong F, Xing J, Li X, Liu X, Fu Z, et al. (2018) Artificial intelligence in drug design. Science China Life Sciences : 1–14.

115. Kubinyi H (2003) Drug research: myths, hype and reality. Nature Reviews Drug Discovery 2: 665.

116. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: applications to targets and beyond. British journal of pharmacology 152: 21–37.

117. Kharkar PS, Warrier S, Gaud RS (2014) Reverse docking: a powerful tool for drug repositioning and drug rescue. Future medicinal chemistry 6: 333–342.

118. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery 3: 711.

119. Yang Y, Adelstein SJ, Kassis AI (2012) Target discovery from data mining approaches. Drug discovery today 17: S16–S23.

120. Zhang X, Crespo A, Fernández A (2008) Turning promiscuous kinase inhibitors into safer drugs. Trends in biotechnology 26: 295–301.

121. Raman K, Yeturu K, Chandra N (2008) targettb: a target identification pipeline for mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. BMC systems biology 2: 109.

122. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, et al. (2011) From in silico target prediction to multi-target drug design: current databases, methods and applications. Journal of proteomics 74: 2554–2574.

123. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, et al. (2009) Predicting new molecular targets for known drugs. Nature 462: 175.

124. Mogire RM, Akala HM, Macharia RW, Juma DW, Cheruiyot AC, et al. (2017) Target-similarity search using plasmodium falciparum proteome identifies approved drugs with anti-malarial activity and their possible targets. PloS one 12: e0186364.

125. Lounkine E, Keiser MJ, Whitebread S, Mikhailov D, Hamon J, et al. (2012) Large-scale prediction and testing of drug activity on side-effect targets. Nature 486: 361.

126. Wu L, Ai N, Liu Y, Wang Y, Fan X (2013) Relating anatomical therapeutic indications by the ensemble similarity of drug sets. Journal of chemical information and modeling 53: 2154–2160.

127. Hopkins AL, Mason JS, Overington JP (2006) Can we rationally design promiscuous drugs? Current opinion in structural biology 16: 127–136.

128. Chen B, Butte A (2016) Leveraging big data to transform target selection and drug discovery. Clinical Pharmacology & Therapeutics 99: 285–297.

129. Kim YA, Wuchty S, Przytycka TM (2011) Identifying causal genes and dysregulated pathways in complex diseases. PLoS computational biology 7: e1001095.

130. H Andrade C, C Silva D, C Braga R (2014) In silico prediction of drug metabolism by p450. Current drug metabolism 15: 514–525.

131. Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT (2016) Computational approaches in target identification and drug discovery. Computational and structural biotechnology journal 14: 177–184.

132. Gobbi G, Janiri L (1999) Clozapine blocks dopamine, 5-ht2 and 5-ht3 responses in the medial prefrontal cortex: an in vivo microiontophoretic study. European Neuropsychopharmacology 10: 43–49.

133. Ashley EA, White NJ (2005) Artemisinin-based combinations. Current opinion in infectious diseases 18: 531–536.

134. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, et al. (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nature communications 8: 573.

135. Harrold J, Ramanathan M, Mager D (2013) Network-based approaches in drug discovery and early development. Clinical Pharmacology & Therapeutics 94: 651–658.

136. Huthmacher C, Hoppe A, Bulik S, Holzhütter HG (2010) Antimalarial drug targets in plasmodium falciparum predicted by stage-specific metabolic network analysis. BMC systems biology 4: 120.

137. Rout S, Patra NP, Mahapatra RK (2017) An in silico strategy for identification of novel drug targets against plasmodium falciparum. Parasitology research 116: 2539–2559.

138. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, et al. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nature biotechnology 30: 159.

139. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug discovery today 23: 1538–1546.

140. Weaver DC (2004) Applying data mining techniques to library design, lead generation and lead optimization. Current opinion in chemical biology 8: 264–270.

141. Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers & chemistry 26: 5–14.

142. Nasrabadi NM (2007) Pattern recognition and machine learning. Journal of electronic imaging 16: 049901.

143. Flach PA, Lachiche N (2004) Naive bayesian classification of structured data. Machine Learning 57: 233–269.

144. Mazandu GK, Opap K, Mulder NJ (2011) Contribution of microarray data to the advancement of knowledge on the mycobacterium tuberculosis interactome: Use of the random partial least squares approach. Infection, Genetics and Evolution 11: 725–733.

145. Fatumo S, Adebiyi E, Schramm G, Eils R, König R (2009) An in silico approach to detect efficient malaria drug targets to combat the malaria resistance problem. In: | 2009 International Association of Computer Science and Information Technology-Spring Conference. IEEE, pp. 576–580.

146. Sturm M, Hackenberg M, Langenberger D, Frishman D (2010) Targetspy: a supervised machine learning approach for microrna target prediction. BMC bioinformatics 11: 292.

147. Nidhi a, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. Journal of chemical information and modeling 46: 1124–1133.

148. Azencott CA, Ksikes A, Swamidass SJ, Chen JH, Ralaivola L, et al. (2007) One-to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. Journal of chemical information and modeling 47: 965–974.

149. Golbraikh A, Wang XS, Zhu H, Tropsha A (2016) Predictive qsar modeling: methods and applications in drug discovery and chemical risk assessment. Handbook of computational chemistry : 1–48.

150. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug discovery today 20: 318–331.

151. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, et al. (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS one 8: e61318.

152. Nigsch F, Bender A, Jenkins JL, Mitchell JB (2008) Ligand-target prediction using winnow and naive bayesian algorithms and the implications of overall performance statistics. Journal of Chemical Information and Modeling 48: 2313–2325.

153. Awale M, Reymond JL (2018) Polypharmacology browser ppb2: Target prediction combining nearest neighbors with machine learning. Journal of chemical information and modeling 59: 10–17.

154. Sperandei S (2014) Understanding logistic regression analysis. Biochemia medica: Biochemia medica 24: 12–18.

155. Lee M, Kim D (2012) Large-scale reverse docking profiles and their applications. In: BMC bioinformatics. BioMed Central, volume 13, p. S6.

156. Sarnpitak P, Mujumdar P, Taylor P, Cross M, Coster MJ, et al. (2015) Panel docking of small-molecule libraries—prospects to improve efficiency of lead compound discovery. Biotechnology advances 33: 941–947.

157. Li H, Gao Z, Kang L, Zhang H, Yang K, et al. (2006) Tarfisdock: a web server for identifying drug targets with docking approach. Nucleic acids research 34: W219–W224.

158. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, et al. (2006) A critical assessment of docking programs and scoring functions. Journal of medicinal chemistry 49: 5912–5931.

159. Lee A, Lee K, Kim D (2016) Using reverse docking for target identification and its applications for drug discovery. Expert opinion on drug discovery 11: 707–715.

160. Byrne R, Schneider G (2019) In silico target prediction for small molecules. In: Systems Chemical Biology, Springer. pp. 273–309.

161. Wang JC, Chu PY, Chen CM, Lin JH (2012) idtarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic acids research 40: W393–W399.

162. Chen Y, Zhi D (2001) Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins: Structure, Function, and Bioinformatics 43: 217–226.

163. Trott O, Olson AJ (2010) Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry 31: 455–461.

164. Ma C, Tang K, Liu Q, Zhu R, Cao Z (2013) Calmodulin as a potential target by which berberine induces cell cycle arrest in human hepatoma b el7402 cells. Chemical biology & drug design 81: 775–783.

165. Scafuri B, Marabotti A, Carbone V, Minasi P, Dotolo S, et al. (2016) A theoretical study on predicted protein targets of apple polyphenols and possible mechanisms of chemoprevention in colorectal cancer. Scientific reports 6: 32516.

166. Wang JC, Lin JH, Chen CM, Perryman AL, Olson AJ (2011) Robust scoring functions for protein–ligand interactions with quantum chemical charge models. Journal of chemical information and modeling 51: 2528–2537.

167. Chang DTH, Lin JH, Hsieh CH, Oyang YJ (2009) On the design of optimization algorithms for prediction of molecular interactions. In: 2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering. IEEE, pp. 208–215.

168. Lauro G, Romano A, Riccio R, Bifulco G (2011) Inverse virtual screening of antitumor targets: pilot study on a small database of natural bioactive compounds. Journal of natural products 74: 1401–1407.

169. Lauro G, Masullo M, Piacente S, Riccio R, Bifulco G (2012) Inverse virtual screening allows the discovery of the biological activity of natural compounds. Bioorganic & medicinal chemistry 20: 3596–3602.

170. Wermuth CG (2006) Selective optimization of side activities: the sosa approach. Drug discovery today 11: 160–164.

171. Wang W, Zhou X, He W, Fan Y, Chen Y, et al. (2012) The interprotein scoring noises in glide docking scores. Proteins: Structure, Function, and Bioinformatics 80: 169–183.

172. Yuriev E, Holien J, Ramsland PA (2015) Improvements, trends, and new ideas in molecular docking: 2012–2013 in review. Journal of Molecular Recognition 28: 581–604.

173. Wang Z, Liang L, Yin Z, Lin J (2016) Improving chemical similarity ensemble approach in target prediction. Journal of cheminformatics 8: 20.

174. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, et al. (2007) Relating protein pharmacology by ligand chemistry. Nature biotechnology 25: 197.

175. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nature chemical biology 4: 682.

176. Fliri AF, Loging WT, Thadeio PF, Volkmann RA (2005) Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. Journal of medicinal chemistry 48: 6918–6925.

177. Jenkins JL, Bender A, Davies JW (2006) In silico target fishing: Predicting biological targets from chemical structure. Drug Discovery Today: Technologies 3: 413–421.

178. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nature biotechnology 24: 805.

179. Fliri AF, Loging WT, Thadeio PF, Volkmann RA (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. Proceedings of the National Academy of Sciences 102: 261–266.

180. Godden JW, Xue L, Bajorath J (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. Journal of Chemical Information and Computer Sciences 40: 163–166.

181. Downs GM, Willett P, Fisanick W (1994) Similarity searching and clustering of chemical-structure databases using molecular property data. Journal of Chemical Information and Computer Sciences 34: 1094–1102.

182. Breu H, Gil J, Kirkpatrick D, Werman M (1995) Linear time euclidean distance transform algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 17: 529–533.

183. de Souza RM, De Carvalho FdA (2004) Clustering of interval data based on city–block distances. Pattern Recognition Letters 25: 353–365.

184. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321: 263–266.

185. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Organic & biomolecular chemistry 2: 3204–3218.

186. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, et al. (2006) Bridging chemical and biological space:"target fishing" using 2d and 3d molecular descriptors. Journal of medicinal chemistry 49: 6802–6810.

187. Raymond JW, Willett P (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2d chemical structure databases. Journal of computer-aided molecular design 16: 59–71.

188. Khan AU, et al. (2016) Descriptors and their selection methods in qsar analysis: paradigm for drug design. Drug Discovery Today 21: 1291–1302.

189. Hert J, sWillett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. Organic & biomolecular chemistry 2: 3256–3266.

190. Gfeller D, Grosdidier A, Wirth M, Daina A, Michielin O, et al. (2014) Swisstargetprediction: a web server for target prediction of bioactive small molecules. Nucleic acids research 42: W32–W38.

191. Reker D, Rodrigues T, Schneider P, Schneider G (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. Proceedings of the National Academy of Sciences : 201320001.

192. Dunkel M, Günther S, Ahmed J, Wittig B, Preissner R (2008) Superpred: drug classification and target prediction. Nucleic acids research 36: W55–W59.

193. Awale M, Reymond JL (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using chembl bioactivity data. Journal of cheminformatics 9: 11.

194. Liu X, Vogt I, Haque T, Campillos M (2013) Hitpick: a web server for hit identification and target prediction of chemical screenings. Bioinformatics 29: 1910–1912.

195. Lagunin A, Stepanchikova A, Filimonov D, Poroikov V (2000) Pass: prediction of activity spectra for biologically active substances. Bioinformatics 16: 747–748.

196. Huang T, Mi H, Lin Cy, Zhao L, Zhong LL, et al. (2017) Most: most-similar ligand based approach to target prediction. BMC bioinformatics 18: 165.

197. Rhodes N, Willett P, Calvet A, Dunbar JB, Humblet C (2003) Clip: similarity searching of 3d databases using clique detection. Journal of chemical information and computer sciences 43: 443–448.

198. Lo YC, Senese S, Li CM, Hu Q, Huang Y, et al. (2015) Large-scale chemical similarity networks for target profiling of compounds identified in cell-based chemical screens. PLoS computational biology 11: e1004153.

199. Cruz-Monteagudo M, Medina-Franco JL, Perez-Castillo Y, Nicolotti O, Cordeiro MND, et al. (2014) Activity cliffs in drug discovery: Dr jekyll or mr hyde? Drug Discovery Today 19: 1069–1080.

200. Vega S, Abian O, Velazquez-Campoy A (2016) On the link between conformational changes, ligand binding and heat capacity. Biochimica et Biophysica Acta (BBA)-General Subjects 1860: 868–878.

201. Schmitt S, Kuhn D, Klebe G (2002) A new method to detect related function among proteins independent of sequence and fold homology. Journal of molecular biology 323: 387–406.

202. Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, et al. (2005) The sumo server: 3d search for protein functional sites. Bioinformatics 21: 3929–3930.

203. Chartier M, Adriansen E, Najmanovich R (2015) Isomif finder: online detection of binding site molecular interaction field similarities. Bioinformatics 32: 621–623.

204. Yeturu K, Chandra N (2008) Pocketmatch: a new algorithm to compare binding sites in protein structures. BMC bioinformatics 9: 543.

205. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d: application to ligand prediction. BMC bioinformatics 11: 99.

206. Huang H, Zhang G, Zhou Y, Lin C, Chen S, et al. (2018) Reverse screening methods to search for the protein targets of chemopreventive compounds. Frontiers in chemistry 6.

207. Wang X, Pan C, Gong J, Liu X, Li H (2016) Enhancing the enrichment of pharmacophore-based target prediction for the polypharmacological profiles of drugs. Journal of chemical information and modeling 56: 1175–1183.

208. Dosa PI, Amin EA (2015) Tactical approaches to interconverting gpcr agonists and antagonists. Journal of medicinal chemistry 59: 810–840.

209. Mervin LH, Cao Q, Barrett IP, Firth MA, Murray D, et al. (2016) Understanding cytotoxicity and cytostaticity in a high-throughput screening collection. ACS chemical biology 11: 3007–3023.

210. Mervin LH, Afzal AM, Brive L, Engkvist O, Bender A (2018) Extending in silico protein target prediction models to include functional effects. Frontiers in Pharmacology 9.

211. Chartier M, Najmanovich R (2015) Detection of binding site molecular interaction field similarities. Journal of chemical information and modeling 55: 1600–1615.

212. Dopazo J (2014) Genomics and transcriptomics in drug discovery. Drug discovery today 19: 126–132.

213. Van't Veer LJ, Bernards R (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452: 564.

214. Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics 13: 523.

215. Fan-Minogue H, Chen B, Sikora-Wohlfeld W, Sirota M, Butte AJ (2014) A systematic assessment of linking gene expression with genetic variants for prioritizing candidate targets. In: Pacific Symposium on Biocomputing Co-Chairs. World Scientific, pp. 383–394.

216. Plenge RM, Scolnick EM, Altshuler D (2013) Validating therapeutic targets through human genetics. Nature reviews Drug discovery 12: 581.

217. Okada Y, Wu D, Trynka G, Raj T, Terao C, et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 506: 376.

218. Hsu YH, Yao J, Chan LC, Wu TJ, Hsu JL, et al. (2014) Definition of pkc-α, cdk6, and met as therapeutic targets in triple-negative breast cancer. Cancer research 74: 4822–4835.

219. Lee HW, Joo KM, Lim JE, Cho HJ, Cho HJ, et al. (2013) Tpl2 kinase impacts tumor growth and metastasis of clear cell renal cell carcinoma. Molecular Cancer Research 11: 1375–1386.

220. Wei L, Liu Y, Dubchak I, Shon J, Park J (2002) Comparative genomics approaches to study organism similarities and differences. Journal of Biomedical Informatics 35: 142–150.

221. Hossain T, Kamruzzaman M, Choudhury TZ, Mahmood HN, Nabi A, et al. (2017) Application of the subtractive genomics and molecular docking analysis for the identification of novel putative drug targets against salmonella enterica subsp. enterica serovar poona. BioMed Research International 2017.

222. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, et al. (2019) Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion 50: 71–91.

223. Wu Z, Li W, Liu G, Tang Y (2018) Network-based methods for prediction of drug-target interactions. Frontiers in pharmacology 9.

224. Hsin KY, Kitano H, Matsuoka Y, Ghosh S (2015) Application of machine leaning approaches in drug target identification and network pharmacology. In: 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). IEEE, pp. 219–219.

225. Lu L, Medo M, Yeung CH, Zhang YC, Zhang ZK, et al. (2012) Recommender systems. CoRR abs/1202.1112.

226. Zhou T (2011) Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390: 1150–1170.

227. Hsin KY, Ghosh S, Kitano H (2013) Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. PloS one 8: e83922.

228. Ballester PJ, Mitchell JB (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26: 1169–1175.

229. Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set. Journal of chemical information and modeling 49: 1079–1093.

230. Plewczynski D, Łaźniewski M, Augustyniak R, Ginalski K (2011) Can we trust docking results? evaluation of seven commonly used programs on pdbbind database. Journal of computational chemistry 32: 742–755.

231. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug discovery today 23: 1241–1250.

232. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, et al. (2018) A primer on deep learning in genomics. Nature genetics : 1.

233. Pan Y, Zhang Y, Liu J (2018) Text mining-based drug discovery in cutaneous squamous cell carcinoma. Oncology reports 40: 3830–3842.

234. Cardon LR, Harris T (2016) Precision medicine, genomics and drug discovery. Human molecular genetics 25: R166–R172.

235. Bari MG, Ung CY, Zhang C, Zhu S, Li H (2017) Machine learning-assisted network inference approach to identify a new class of genes that coordinate the functionality of cancer networks. Scientific reports 7: 6993.

236. Wale N, Karypis G (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. Journal of chemical information and modeling 49: 2190–2201.

237. Csermely P, Agoston V, Pongor S (2005) The efficiency of multi-target drugs: the network approach might help drug design. Trends in pharmacological sciences 26: 178–182.

238. Li P, Huang C, Fu Y, Wang J, Wu Z, et al. (2015) Large-scale exploration and analysis of drug combinations. Bioinformatics 31: 2007–2016.

239. Zhao XM, Iskar M, Zeller G, Kuhn M, Van Noort V, et al. (2011) Prediction of drug combinations by integrating molecular and pharmacological data. PLoS computational biology 7: e1002323.

240. Altevogt BM, Davis M, Pankevich DE, Norris SMP, et al. (2014) Improving and Accelerating Therapeutic Development for Nervous System Disorders: Workshop Summary. National Academies Press.

241. Magalingam KB, Radhakrishnan A, Ping NS, Haleagrahara N (2018) Current concepts of neurodegenerative mechanisms in alzheimer's disease. BioMed research international 2018.

242. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, et al. (2014) Drug resistance in cancer: an overview. Cancers 6: 1769–1792.

243. Vasilakou E, Machado D, Theorell A, Rocha I, Nöh K, et al. (2016) Current state and challenges for dynamic metabolic modeling. Current opinion in microbiology 33: 97–104.

244. Kent WJ (2002) Blat—the blast-like alignment tool. Genome research 12: 656–664.

245. Mazandu GK, Mulder NJ (2011) Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One 6: e18607.

246. De Las Rivas J, Fontanillo C (2010) Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS computational biology 6: e1000807.

247. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417: 399.

248. Mazandu GK, Mulder NJ (2011) Generation and analysis of large-scale data-driven mycobacterium tuberculosis functional networks for drug target identification. Advances in bioinformatics 2011.

249. Pitre S, Alamgir M, Green JR, Dumontier M, Dehne F, et al. (2008) Computational methods for predicting protein–protein interactions. In: Protein–Protein Interaction, Springer. pp. 247–267.

250. Pearson WR (2013) An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics 42: 3–1.

251. Bastien O, Ortet P, Roy S, Maréchal E (2005) A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise z-score probabilities. BMC bioinformatics 6: 49.

252. Mulder NJ, Akinola RO, Mazandu GK, Rapanoel H (2014) Using biological networks to improve our understanding of infectious diseases. Computational and structural biotechnology journal 11: 1–10.

253. Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

254. Ismail HM, Barton V, Phanchana M, Charoensutthivarakul S, Wong MH, et al. (2016) Artemisinin activity-based probes identify multiple molecular targets within the asexual stage of the malaria parasites plasmodium falciparum 3d7. Proceedings of the National Academy of Sciences 113: 2080–2085.

255. Cheeseman IH, Miller BA, Nair S, Nkhoma S, Tan A, et al. (2012) A major genome region underlying artemisinin resistance in malaria. science 336: 79–82.

256. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proceedings of the national academy of sciences 99: 7821–7826.

257. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Physical review E 69: 026113.

258. Pons P, Latapy M (2005) Computing communities in large networks using random walks. In: International symposium on computer and information sciences. Springer, pp. 284–293.

259. Pons P, Latapy M (2006) Computing communities in large networks using random walks. J Graph Algorithms Appl 10: 191–218.

260. Wu F, Huberman BA (2004) Finding communities in linear time: a physics approach. The European Physical Journal B 38: 331–338.

261. Ponts N, Yang J, Chung DWD, Prudhomme J, Girke T, et al. (2008) Deciphering the ubiquitin-mediated pathway in apicomplexan parasites: a potential strategy to interfere with parasite virulence. PloS one 3: e2386.

262. Hamilton MJ, Lee M, Le Roch KG (2014) The ubiquitin system: an essential component to unlocking the secrets of malaria parasite biology. Molecular bioSystems 10: 715–723.

263. Sharma M, Dhiman C, Dangi P, Singh S (2014) Designing synthetic drugs against plasmodium falciparum: a computational study of histone-lysine n-methyltransferase (pfhkmt). Systems and synthetic biology 8: 155–160.

264. Villard V, Agak GW, Frank G, Jafarshad A, Servis C, et al. (2007) Rapid identification of malaria vaccine candidates based on α-helical coiled coil protein motif. PloS one 2: e645.

265. Cui L, Fan Q, Cui L, Miao J (2008) Histone lysine methyltransferases and demethylases in plasmodium falciparum. International journal for parasitology 38: 1083–1097.

266. Kaur I, Zeeshan M, Saini E, Kaushik A, Mohmmed A, et al. (2016) Widespread occurrence of lysine methylation in plasmodium falciparum proteins at asexual blood stages. Scientific reports 6: 35432.

267. Ludin P, Woodcroft B, Ralph SA, Mäser P (2012) In silico prediction of antimalarial drug target candidates. International journal for parasitology: drugs and drug resistance 2: 191–199.

268. Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbsnp: a database of single nucleotide polymorphisms. Nucleic acids research 28: 352–355.

269. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, et al. (2018) Genemania update 2018. Nucleic acids research 46: W60–W64.

270. Pleass RJ, Holder AA (2005) Antibody-based therapies for malaria. Nature Reviews Microbiology 3: 893.

271. Tripathi AK, Sha W, Shulaev V, Stins MF, Sullivan DJ (2009) Plasmodium falciparum–infected erythrocytes induce nf-κb regulated inflammatory pathways in human cerebral endothelium. Blood 114: 4243–4252.

272. Lyke KE, Fernández-Viña MA, Cao K, Hollenbach J, Coulibaly D, et al. (2011) Association of hla alleles with plasmodium falciparum severity in malian children. Tissue Antigens 77: 562–571.

273. Guo H, Callaway JB, Ting JP (2015) Inflammasomes: mechanism of action, role in disease, and therapeutics. Nature medicine 21: 677.

274. Cheng C, Ho WE, Goh FY, Guan SP, Kong LR, et al. (2011) Anti-malarial drug artesunate attenuates experimental allergic asthma via inhibition of the phosphoinositide 3-kinase/akt pathway. PLoS One 6: e20932.

275. Hooper LV, Stappenbeck TS, Hong CV, Gordon JI (2003) Angiogenins: a new class of microbicidal proteins involved in innate immunity. Nature immunology 4: 269.

276. Bridgford JL, Xie SC, Cobbold SA, Pasaje CFA, Herrmann S, et al. (2018) Artemisinin kills malaria parasites by damaging proteins and inhibiting the proteasome. Nature communications 9: 3801.

277. Chen LC, Yeh HY, Yeh CY, Arias CR, Soo VW (2012) Identifying co-targets to fight drug resistance based on a random walk model. BMC systems biology 6: 5.

278. Eriksson E, Sampaio N, Schofield L (2013) Toll-like receptors and malaria–sensing and susceptibility. J Trop Dis 2.

279. Greene JA, Moormann AM, Vulule J, Bockarie MJ, Zimmerman PA, et al. (2009) Toll-like receptor polymorphisms in malaria-endemic populations. Malaria journal 8: 50.

280. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, et al. (2014) Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic acids research 43: D1071–D1078.

281. Mathur S, Dinakarpandian D (2012) Finding disease similarity based on implicit semantic similarity. Journal of biomedical informatics 45: 363–371.

282. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, et al. (2016) Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research : gkw943.

283. Mazandu GK, Chimusa ER, Mulder NJ Expressions of different semantic similarity measures in the context of biomedical sciences and wordnet .

284. Clark IA, Alleva LM, Mills AC, Cowden WB (2004) Pathogenesis of malaria and clinically similar conditions. Clinical microbiology reviews 17: 509–539.

285. Murphy SC, Breman JG (2001) Gaps in the childhood malaria burden in africa: cerebral malaria, neurological sequelae, anemia, respiratory distress, hypoglycemia, and complications of pregnancy. The American journal of tropical medicine and hygiene 64: 57–67.

286. Dunst J, Kamena F, Matuschewski K (2017) Cytokines and chemokines in cerebral malaria pathogenesis. Frontiers in cellular and infection microbiology 7: 324.

287. Kumar R, Ng S, Engwerda C (2019) The role of il-10 in malaria: a double edged sword. Frontiers in immunology 10: 229.

288. Franklin BS, Ishizaka ST, Lamphier M, Gusovsky F, Hansen H, et al. (2011) Therapeutical targeting of nucleic acid-sensing toll-like receptors prevents experimental cerebral malaria. Proceedings of the National Academy of Sciences 108: 3689–3694.