GENETIC EPIDEMIOLOGY ALGORITHMS FOR TRACKING DRUG RESISTANCE VARIANTS AND GENOMIC CLUSTERING OF *PLASMODIUM* SPECIES

A Dissertation
Presented to
The Academic Faculty

by

Shashidhar Ravishankar

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology
December 2019

GENETIC EPIDEMIOLOGY ALGORTIHMS FOR TRACKING DRUG RESISTANCE VARIANTS AND GENOMIC CLUSTERING OF *PLASMODIUM* SPECIES

Approved by:

Dr. Fredrik O. Vannberg, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Venkatachalam Udhayakumar
Malaria Branch
Centers for Disease Control and Prevention

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. John McDonald
School of Biological Sciences
Georgia Institute of Technology

Dr. Eberhard Voit
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology

Date Approved: August 21, 2019

If I have seen further, it is by standing on the shoulders of giants

Issac Newton

To my family,

Appa, Amma, Mahesh, Haritha and Ananya

ACKNOWLEDGEMENTS

I'd like to thank my committee for taking time off their busy schedule to offer their advice and guidance throughout this process.

Dr. King Jordan whose support and guidance, from my time as a master's student and throughout my Ph.D., has been invaluable. His advice has been key in shaping the work presented here in my thesis.

Dr. Udhayakumar Venkatachalam is one of the most patient and helpful people I have come across. His support has been critical to everything I have done throughout my Ph.D. The work done in this Ph.D. was made possible through the collaborations with the Dr. Kumar and the Malaria Branch. I will always cherish the time I got to work with him and the Malaria Branch at CDC. I'm grateful for his unwavering support throughout my time as a Ph.D. student.

Dr. Eberhard Voit brings a biological perspective to my committee. Much of my work has focused on Malaria genomics. His experience working with the MAHPIC consortium brings in a perspective that I needed in my committee meetings.

Dr. John McDonald was the first collaborator I worked with after starting at the Vannberg Lab as a master's student. The work done in collaboration with the McDonald Lab laid the foundation for everything I did during the Ph.D.

Dr. Fredrik Vannberg has been the most supportive mentor and advisor. His guidance has been pivotal in everything I have been able to achieve as a Ph.D. student. His unwavering support of everything I have done at the lab is something I will never forget. His passion for bioinformatics and the eagerness with which he approaches science is something I hope to emulate wherever I go next.

I would also like to thank Dr. Peter Audano, Dr. Cai Huang, and Kizee Etienne. Brainstorming with them was about everything bioinformatics was the highlight of my time at the Vannberg Lab. Their advice helped get through many hurdles that I faced as a graduate student.

My work on exploring the epidemiology of malaria was inspired by Dr. Eldin Talundzic. Eldin's passion towards applying bioinformatics in malaria epidemiology has been an inspiring force. He ever-exploring attitude has kept me motivated to explore new ideas throughout my PhD. I would also like to thank Dhruviben Patel, Dr. Sarah Schmedes, Julia Kelly, Dr. Joel Barratt and everyone on the MaRS team, who made my time as an intern at the CDC an unforgettable experience.

My time at Georgia Tech would not have been as enjoyable as it was, if it had not been for the amazing people I met along the way. I would especially like to thank Vaishnavi Venkat, Shruti Lall, Anish Mukherjee and Anusha Harish, for being my second family while I was at Tech. Their friendship will stay with me for the rest of my life. Vaishnavi Venkat also illustrated Figure 1.1.

Most importantly I would like to thank my family, my parents for always being supportive and encouraging me to pursue my interests at every stage of my life. My brother and sister-in-law for their unwavering support and encouragement. Finally, my niece Ananya for brightening up my day, every day for the past two years.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iv**

LIST OF TABLES **viii**

LIST OF FIGURES **ix**

SUMMARY **xii**

CHAPTER 1. An introduction to genomics in malaria epidemiology **1**
1.1 Abstract **1**
1.2 Malaria parasites and their life cycle **2**
1.3 Burden of Malaria **5**
1.4 Vector Control **7**
1.5 Malaria vaccines **9**
1.6 Case management methodologies for malaria **9**
1.6.1 Diagnosis of Malaria 9
1.6.2 Drug treatment 10
1.6.3 Monitoring drug resistance in malaria 11
1.6.4 Molecular markers of drug resistance 12
1.7 Progression from Sanger sequencing to Next Generation Sequencing (NGS) for molecular surveillance **16**
1.8 Next Generation Sequencing (NGS) solutions for malaria epidemiology **17**
1.9 Algorithms to monitor outbreaks and molecular surveillance of drug resistance using Next Generation Sequencing (NGS). **20**

CHAPTER 2. Variant calling and Genomic clustering **24**
2.1 Abstract **24**
2.2 Variant calling from Next Generation Sequencing (NGS) data **25**
2.2.1 Steps involved in variant calling 27
2.2.2 Variant calling 30
2.2.3 Consensus-based variant calling 32
2.3 Genomic clustering **34**
2.3.1 Alignment based genomic clustering 34
2.3.2 SNP based genomic clustering 35
2.3.3 k-mer based genomic clustering 35

CHAPTER 3. k-mer based clustering algorithms to identify relatedness of species from whole genome seuqencing data **38**
3.1 Abstract **38**
3.2 Introduction **39**
3.3 k-mer based clustering algorithm to identify the relatedness of species from NGS data **41**

3.4 Evaluating the accuracy of Gentoo in comparison to ANI as a methodology to identify relatedness of *Plasmodium* spp. 45
3.4.1 Materials and methods 45
3.4.2 Results 48
3.5 Clustering outbreaks of *Candida auris* infections in Colombia 55
3.5.1 Materials and methods 55
3.5.2 Results 56
3.6 Discussion 57

CHAPTER 4. Next Generation Sequencing and bioinformatics protocol for malaria drug resistance marker surveillance 60
4.1 Abstract 60
4.2 Introduction 61
4.3 NeST variant calling framework 63
4.3.1 PrepInputs Module 63
4.3.2 VarCallEngine Module 64
4.3.3 VCFToolkit Module 66
4.3.4 Summarize Module 66
4.3.5 Accessibility and cross platform compatibility 67
4.3.6 Input and Result standardization 70
4.4 *In-silico* evaluation of variant calling accuracy from NGS datasets 74
4.4.1 Materials and methods: 74
4.4.2 Results: 78
4.5 Identifying variants conferring antimalarial drug resistance in *Plasmodium falciparum* from Targeted Amplicon Deep Sequencing datasets 80
4.5.1 Materials and methods 80
4.5.2 Results 82
***4.6* Evaluating accuracy of variant calling derived genotypes at predicting phenotypic drug resistance in *Mycobacterium tuberculosis* 87**
4.6.1 Material and methods 87
4.6.2 Results: 88
4.7 Discussion 92

CHAPTER 5. Concluding remarks 94

REFERENCES 97

LIST OF TABLES

Table 1.1	Common antimalarial drug and genetic markers associated with drug resistance in *P. falciparum*. *crt*, chloroquine-resistance transporter; *cytb*, cytochrome b; *dhfr*, dihydrofolate reductase; *dhps*, dihydropteroate synthase; *mdr1*, multidrug resistance protein; pfkelch*13*, *P. falciparum* Kelch 13; *plm2*, plasmepsin 2. * Drug used in artemisinin-based combination therapy; Ψ Antimalarial drug used alone or in combination with molecules other than artemisinin derivatives.	13
Table 1.2	Candidate and validated resistance mutations in the *K13* BTB/POZ and propeller domain.	15
Table 3.1	*In-silico* datasets generated from each of the *Plasmodium* spp., genomes.	47
Table 4.1	Comparison of usability features available across open-source variant calling platforms.	68
Table 4.2	Variants of interest table. Each row should contain the Chromosome, Gene name, reference amino acid, alternate amino acid and the amino acid location for the variant of interest.	72
Table 4.3	List of summary files generated by NeST.	72
Table 4.4	Characteristics of *in-silico* amplicon data set generated.	77
Table 4.5	Precision and recall values for SNPs and InDel calls made by standard variant callers against *in-silico* datasets from *Plasmodium falciparum* genes.	78
Table 4.6	Truth table for genotypic and phenotypic resistance or susceptibility.	90
Table 4.7	Precision and recall values for the genotypic prediction of phenotypic resistance or susceptibility against anti-TB drugs.	91

LIST OF FIGURES

Figure 1.1 Stages of *Plasmodium* life cycle. Human infection begins with the delivery of sporozoites by the bite of an infected female Anopheles mosquito during a blood meal. These sporozoites migrate to the liver, via the bloodstream and infect hepatocytes. Following a phase of asexual replication, they develop into merozoites. Merozoites are released into the blood and invade the red blood cells. The parasite forms the ring stage and subsequently develops into trophozoite and schizont stages. The schizont burst to release more merozoites, and this forms the asexual blood stages of the parasite. A small percentage of the ring form parasites differentiate into gametocytes and taken up by the mosquito during a blood meal. The gametocytes travel to the midgut of the mosquito and undergo sexual reproduction to form a zygote, which matures into an oocyst. The oocysts burst to release sporozoites that travel to the salivary gland of the mosquito ready to infect the next host. 4

Figure 1.2 Countries with indigenous cases of malaria in 2000 and their status in 2017. Countries that reported no indigenous cases in the past 3 consecutive years are classified as malaria free. All WHO countries in the European Region report zero indigenous cases in 2016 and 2017. China and El Salvador report no indigenous cases in 2017. Source: World Malaria Report 2018. 8

Figure 1.3 The pathways involved in the action of antimalarial drugs and the molecular markers that affect the resistance. Source: from *Blasco* et al. 2017. 14

Figure 1.4 Emergence and spread of *P. falciparum* resistance to chloroquine, pyrimethamine, and artemisinin derivatives. Resistance to chloroquine emerged at multiple sites and spread across the world (black arrows), due to the selective pressure on *PfCRT* mutant alleles. Resistance to pyrimethamine emerged in South East Asia and South America. Resistance to pyrimethamine due to a triple mutation in *PfDHFR* spread to Africa (red arrows). Pyrimethamine-resistant *PfDHFR* mutations independently emerged in Africa. Resistance to artemisinin derivatives were driven by mutant *PfK13* alleles and were first detected in South East Asia. Source: Blasco et al., 2017. 18

Figure 1.5 MaRS protocol workflow overview with steps indicated, along with reagents needed, total time, and hands-on time for each procedure. 23

Figure 2.1 Cost per MB of DNA sequenced. To highlight the improvements in sequencing technology, the graph shows a line reflecting Moore's Law, which states that "compute power" doubles every two years. Technologies that keep up with Moore's Law are considered to be doing exceedingly well. The y-axis uses a logarithmic scale. The sudden drop in the cost of 24

sequencing in 2008, corresponds to the switch from Sanger-based to second-generation sequencing technologies.

Figure 3.1 Overview of the steps involved in k-mer counting. a) Raw DNA sequences from FASTA or FASTQ files. b) k-mer count tables at 4-mers for the sequence from (a) in string representation. c) 2-bit encoding scheme for DNA sequences. d) Integer representation of k-mer count tables. d) Conversion of string representation of k-mer to integer representation. 44

Figure 3.2 Comparison of neighbor-joining trees generated from pairwise distance estimation made using Gentoo (3.2b, e), Mash (3.2c, f), ANI (3.2d), and Finch (3.2g). Accuracy of the branch points was determined using the *Plasmodium* evolutionary tree (3.2a) published by Rutledge et al., 2017 Branching point comparison shows that the neighbor-joining tree generated by Gentoo is the closest to the *Plasmodium* evolutionary tree, with the exception being the branch point for *P. ovale curtisi*. This hold true even when estimating distances from in-silico simulated FASTQ files (3.2e). Tree generated using ANI, Mash and Finch showed at least two branching point errors. 50-53

Figure 3.3 Memory utilization by ANIm, Mash, Finch and Gentoo as a functional of time. Memory used by each tool for the clustering of 20 *Plasmodium* genomes was recorded. ANI, Mash and Gentoo were run using 30 concurrent processes. Finch does not allow for the user to specify number of concurrent processes. 54

Figure 3.4 Tree generated from *C. auris* outbreak isolates from Colombian and Venezuelan isolates (BioProject ID: PRJNA470683). Clades are shaded based on the geographical locations from which the samples within the clade were isolated. Samples shaded in red were isolated from Hospital A in Cartagena, those shaded in blue were isolated from Hospital B in Barranquilla, and those in purple were isolated from Hospital D in Bogota. a) Maximum parsimony tree of *C. auris* isolates. b) Neighbor-joining tree generated using Gentoo. 58

Figure 4.1 NeST flowchart detailing the four main modules. PrepInputs consolidates all user inputs. VarCallEngine executes three variant calling pipelines. VCFToolkit annotates the variant calls and merges VCF files to provide a consensus variant call. Summarizer generates human readable reports and figures from the NGS analysis. 65

Figure 4.2 NeST virtual environment. 69

Figure 4.3 Depth of sequencing coverage across regions of interest. The y-axis indicates the sequencing read depth; the x-axis lists the variants of interest as specified by the user. Here, the variants of interest are mutations conferring antimalarial drug resistance. 75

Figure 4.4 Allele frequency distribution across variant of interest. The y-axis lists variants of interest as specified by the user. The x-axis indicates the allele frequency in the sample set. The color of the bars indicates the allele 76

balance for the variant. Blue indicates wild type, green indicates variant in the minor allele, and red indicates presence of major allele variant.

Figure 4.5 Distribution of false positive and false negative calls made by the different variant calling methods as a function of sequencing depth and error rates. The swarm plot shows the density of false positive and false negative calls for each condition. 79

Figure 4.6 Summary of the geographical location of all 243 *P. falciparum* samples. Legend indicates the number of samples from each region. 82

Figure 4.7 The overlap of variant calls made by NeST, Geneious, and Sanger sequencing calls. The bar graph on the left shows the total number of variant calls made by the different methods. The graph on the right shows the extent of overlap of variant calls made by NeST, Geneious, and Sanger calls. 84

Figure 4.8 The overlap of variant calls made by NeST, Geneious, and Sanger sequencing calls. The bar graph on the left shows the total number of variant calls made by the different methods. The graph on the right shows the extent of overlap of variant calls made by Samtools, GATK HaplotypeCaller, Freebayes, Geneious, and Sanger calls. Vertical bars are colored by the methods that identified the variant. 85

Figure 4.9 Frequency of novel non-synonymous mutations in the exonic region PfK13 that were found by at least two variant callers. 86

Figure 4.10 NeST/ MaRS app available on the OAMD Bioinformatics Platform. 87

SUMMARY

Malaria is endemic in many parts of the world, including regions of sub-Saharan Africa, South America, and parts of Asia. Currently, there are five species known to cause malaria in humans: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and *P. knowlesi*. Among them, *P. knowlesi* is a zoonotic parasite restricted to mostly South East Asia. According to the World Malaria Report from 2018 [1], these five species were responsible for nearly 219 million infections, resulting in an estimated 435,000 deaths related to malaria in 2017. One of the milestones set by the World Health Organization (WHO) Global Technical Strategy is the elimination of malaria in at least ten countries that were malaria endemic in 2015[1].

Malaria control strategies still rely on traditional vector control efforts, such as indoor residual spraying and the use of insecticide-treated bednets along with case management, using proper diagnosis and treatment. However, there has been an increase in the number of malaria infections that are resistant to first-line therapies, such as artemisinin combination therapies (ACT) [1]. Reports of resistance are especially high in parts of South East Asia [1]. In order to combat the spread of resistance, it is essential to monitor the movement of drug-resistant parasites using molecular markers of resistance and adopt appropriate treatment strategies. Recent progress in genomics research has helped to identify genetic markers associated with resistance and use them to understand the molecular mechanisms involved in the development of resistance, study the evolutionary dynamics of resistance, and track the spread of resistance.

The increasing throughput and decreasing costs of Next Generation Sequencing (NGS) technologies enable large-scale surveillance of outbreaks and spread of drug

resistance. However, there is a lack of standardized tools for the analysis of the large amounts of data produced from NGS sequencers, especially in a public health context. The work presented in this thesis describes solutions for the surveillance of outbreaks of infectious diseases and the spread of drug resistance on a global scale. The goal of this dissertation is to develop fast and efficient algorithms that can be scaled according to the computational resources available.

Genomic clustering is a commonly used technique to understand the relatedness of isolates from outbreaks of infectious diseases. Traditional methods used to estimate genetic distances between organisms use pairwise or multiple sequence alignment to identify regions of similarity. Significant advances have been made in alignment-based genomic clustering techniques over the past decade, however, they still rely on computationally intensive processes. In an effort to overcome this challenge, I explored the use of alignment-free methods for genomic clustering, which often improves the performance of an algorithm by an order of magnitude.

Alignment free methods usually start with breaking genomic sequences into overlapping fragments of the same length, called k-mers. For example, if the size of the fragment (k) is 31, we would identify and store all substrings of length 31 from the genomic data. Each substring and its corresponding frequency in the genome can be used as a proxy for genomic diversity. The method I propose in Chapter 3, Gentoo, uses the overlap of k-mer sets from sequencing read data of two organisms to estimate their relatedness. This information can be used to cluster isolates in an outbreak. This new method provides a more scalable and accurate platform for genomic clustering as compared to other established techniques and can significantly impacts genetic epidemiology in public health.

Working with the Malaria Branch at the Centers for Disease Control and Prevention (CDC), I realized the need for more robust methods to track the spread of drug resistance. The development of drug resistance in a pathogen is usually a result of the selective pressure due to treatment. Ineffective or incomplete treatment and control strategies end up selecting pathogens that have developed mutations conferring resistance to that drug treatment. Drug resistance can severely affect the epidemiological effort to contain the spread of infectious diseases such as malaria.

Through treatment efficacy studies (TES) and genome-wide association studies (GWAS), molecular markers associated with resistance to antimalarial drugs have been identified. These markers are generally Single Nucleotide Polymorphisms (SNPs) in essential genes involved with binding or efflux of drug molecules. Considering this, I developed a framework to identify the prevalence of SNPs associated with drug resistance from NGS data. This framework was adopted by The Malaria Branch at CDC for the surveillance of antimalarial drug resistance.

Identifying SNPs from sequencing reads involves aligning the reads against a reference genome and calling SNPs from the alignment using variant calling algorithms. In the past decade, many different algorithms have been developed for the identification of SNPs from sequencing data. However, recent publications have shown that variant calls made by different algorithms on the same dataset are not always the same. Relying on a single variant calling methodology makes it hard to distinguish between true variants and sequencing errors. Commonly used variant call filtering techniques rely on organism-specific, population-scale databases to identify true variant calls. Population level information, however, is not available for all organisms. To address this concern, in

Chapter 4, I describe a consensus-based variant calling framework called NeST, which implements multiple standard variant calling pipelines and generates a consensus variant call. NeST provides a framework enabling the identification of high confidence variant calls and overcomes the inherent biases of the statistical models implemented in existing methodologies. The consensus framework also provides a metric to identify true variants from sequencing reads when standard variant filtration techniques cannot be utilized.

NeST implements a scalable consensus variant calling framework that accurately identifies high confidence variant calls associated with drug resistance. NeST forms the foundation for the development of a surveillance system to track the global spread of drug resistance in malaria at the CDC.

The following chapters will describe the current state-of-the-art and the novel improvements developed through this research towards improving genetic epidemiology in public health.

CHAPTER 1. AN INTRODUCTION TO GENOMICS IN MALARIA EPIDEMIOLOGY

1.1 Abstract

Malaria is endemic in many parts of the world, including regions of sub-Saharan Africa, South America, South Asia, and South East Asia. Currently, five known species cause malaria in humans: *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and *P. knowlesi*. Among them, *P. knowlesi* is a zoonotic malaria parasite with transmission localized to South East Asia. According to a World Health Organization (WHO) report from 2018 [1], these five species were responsible for nearly 219 million infections, resulting in an estimated 435,000 deaths related to malaria in 2017. One of the 2020 milestones for the WHO Global technical strategy for malaria 2016-2030, is the elimination of malaria in at least ten countries that were malaria endemic in 2015 [1]. The development of novel molecular tools that can improve the detection of various *Plasmodium* species and monitor the spread of drug resistance in *P. falciparum*, will help to improve surveillance, and facilitate malaria control and elimination goals. The advances of Next Generation Sequencing provides a cost-effective solution for large-scale surveillance of drug resistance using molecular markers of resistance. However, there is a need for standardized, scalable frameworks for analysis of the vast amount of information generated from NGS studies. The methods described in this work propose two novel algorithms for the accurate and scalable analysis of genomic data in a public health setting.

1.2 Malaria parasites and their life cycle

Malaria is a disease caused by infections from parasites of the *Plasmodium* genus. Five species of the *Plasmodium* genus are known to cause malaria in humans, namely *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale* and *Plasmodium knowlesi*. Malaria due to *P. knowlesi* is a zoonotic form of malaria and is known to be an important contributor to malaria in humans living in South East Asia.

Being a vector borne disease, malaria infections spread through bites from female *Anopheline* mosquitoes. The sporozoite is released into the human blood stream when the female *Anopheles* mosquitoes take a blood meal. The sporozoites then make their way to the liver, where they infect the hepatocytes. The transport of the sporozoites to the liver takes about 1-3 hours. The sporozoites that fail to enter the bloodstream are destroyed by the host immune system [2]. Once the sporozoites have infected the hepatocytes, they start to divide mitotically and transform into a schizont. This process takes about 2-10 days. At the end of the liver stage, up to 40,000 merozoites per infected hepatocyte are released into the bloodstream.

In cases of malaria caused by *P. vivax* and *P. ovale* infections, some parasites enter a dormant phase during the liver stage of infection to form a hypnozoite. They can remain in this dormant stage for years. Thus, for infections caused by *P. vivax* and *P. ovale*, there can be a recrudescence of the disease years after the first infection.

The release of the merozoites into the bloodstream marks the start of the asexual blood stage of the parasite life cycle. These merozoites quickly invade erythrocytes. Upon invading the erythrocyte, merozoites undergo several rounds of asexual reproduction to

form 16-32 merozoites over a period of 48 hours, especially with *P. falciparum* and *P. vivax* infections. The maturation cycle varies from 24 hours (*P. knowlesi*) to 72 hours (*P. malariae*). When the mature schizonts rupture, newly formed merozoites are released to infect more red blood cells.

Some of the asexual blood-stage parasites differentiate into the male and female sexual stages of the parasite development called gametocytes. The gametocytes have varying times of maturation, based on the species of *Plasmodium* involved. In *P. falciparum* infections, the gametocyte takes about 8-10 days for maturation. The mosquito ingests gametocytes during the blood meal. The ingestion of the gametocytes marks the beginning of the sexual reproduction stage of the parasite's development in the vector.

Male and female gametocytes fuse and undergo fertilization to form a diploid zygote. The zygote then develops into ookinetes and then into oocysts in the mosquito midgut. Oocysts maturation again varies in time between different species of *Plasmodium* family. In *P. falciparum*, this period spans about 11-16 days, at which point the oocysts burst, releasing infectious sporozoites that travel to the salivary gland of the mosquito. Sporozoites from the salivary glands can re-infect human hosts during a blood meal and thus continuing the cycle of infection [2–4].

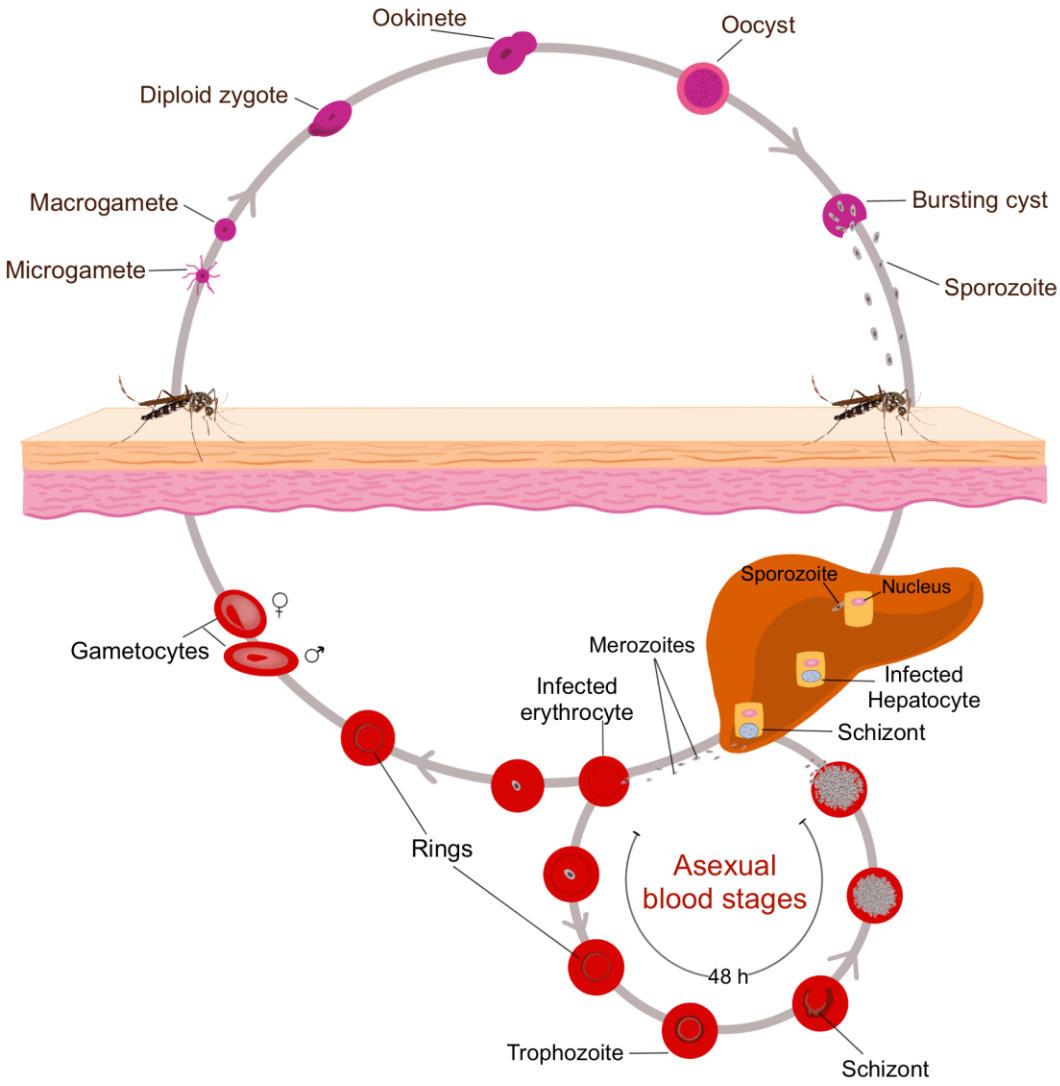


Figure 1.1: Stages of the *Plasmodium* life cycle. Human infection begins with the delivery of sporozoites by the bite of an infected female *Anopheles* mosquito during a blood meal. These sporozoites migrate to the liver, via the bloodstream and infect hepatocytes. Following a phase of asexual replication, they develop into merozoites. Merozoites are released into the blood and invade the red blood cells. The parasite forms the ring stage and subsequently develops into trophozoite and schizont stages. The schizont burst to release more merozoites, and this forms the asexual blood stages of the parasite. A small percentage of the ring form parasites differentiate into gametocytes and taken up by the mosquito during a blood meal. The gametocytes travel to the midgut of the mosquito and undergo sexual reproduction to form a zygote, which matures into an oocyst. The oocysts burst to release sporozoites that travel to the salivary gland of the mosquito ready to infect the next host.

1.3 Burden of Malaria

The World Malaria Report of 2018 estimated that there are 219 million annual cases of malaria worldwide. Though there was a reduction in the number of cases of malaria reported in 2017, as compared to 2010, data between 2015 and 2017 show that there has not been a significant reduction in the number of cases reported worldwide [1].

The majority of cases reported in 2017 were from the African region (92% of reported cases), with the South East Asia region accounting for 5% of the cases and the Eastern Mediterranean region accounting for about 2% of the cases. Nearly half of the malaria burden in the world is borne by five countries, namely, Nigeria, the Democratic Republic of the Congo, Mozambique, India, and Uganda. While India saw a 24% decrease in the rate of incidence over since 2016, Nigeria, Madagascar and the Democratic Republic of the Congo, all reported an increase in the number of cases reported by over half a million cases across each country.

Transmission rates of malaria vary across the world due to many different factors. Transmission intensity ranges from a few infectious bites/year (low transmission areas) to several hundred infectious bites/year (holoendemic areas). The transmission rates play a crucial role in the manifestation of the disease and the development of natural immunity. In regions with high transmission, immunity to clinical malaria develops typically in the first five years of life. The most common manifestations of severe malaria include cerebral malaria, severe malarial anemia and respiratory distress in high endemic areas [2].

In regions of low transmission rates, such as South East Asia, severe malaria is not restricted to children. In low transmission areas, the development of host immunity takes a

long time (until adulthood), and therefore severe malaria can be observed in children and adults. In addition to cerebral malaria and severe malarial anemia, malaria complications may include renal and hepatic dysfunction and multiple system disorders contributing to death[5].

Furthermore, pregnant women are more vulnerable to malaria than their non-pregnant counterparts. The principal manifestation of malaria during pregnancy can vary depending upon the pre-existing immunity. In low transmission regions with limited immunity against malaria, pregnant women can experience severe consequences due to malaria, including cerebral malaria, severe anemia, hypoglycemia, abortion, and stillbirth. On the other hand, in areas with high transmission of malaria, pregnant women rarely experience the severe complication of malaria due to pre-existing immunity. However, in this region malaria infection in pregnant women leads to the development of anemia and low birth weight babies[2].

The number of deaths due to malaria have decreased from 451,000 deaths in 2016 to 435,000 deaths in 2017. It is important to note that young children (under the age of 5) especially in Africa, remain a major susceptible group accounting for 266,000 deaths (61%) associated with malaria[1].

As there are no highly efficacious vaccines against malaria, vector control and case management, through accurate diagnosis of malaria and drug treatment, remain major pillars of malaria control strategy as discussed in the next section.

1.4 Vector Control

Vector control mainly focuses on the mosquito, *Anopheles* species, that are involved in the spread of the parasite. The most widely adopted method for vector control across the world has been Insecticide Treated Nets (ITN). According to WHO, half of the population at risk of malaria infections had access to ITNs in the year 2017. Another standard method for vector control is Indoor Residual Spraying (IRS), which involves spraying insecticides indoors [1]. Studies have shown that both of these preventive methods have been highly effective at reducing the risk of malaria infections in regions with endemic malaria [6,7].

With the advent of gene-editing technologies such as CRISPR-Cas9, researchers have also been working on introducing "gene drives" into mosquitoes to prevent the proliferation of the vector, and in turn, control the spread of malaria. In brief, these methods rely on identifying regions in the mosquito genome that confer sterility; a CRISPR-Cas9 gene drive is introduced in these regions to modify the gene conferring sterility in the mosquito[8,9]. Another candidate target for the gene drive is the mechanisms that allow for the parasite to complete its life cycle within the mosquito[10]. The mechanism to modify the gene associated with sterility or parasite uptake is built into the mosquito genome through the gene drive; this method can ensure that these genetic modification are carried across many generations, with the potential to cause the extinction of the species capable of harboring the malaria parasite[10,11]. While this approach provides an interesting mechanism for vector control, the efficacy and ethics involved in vector control are beyond the scope of this thesis.

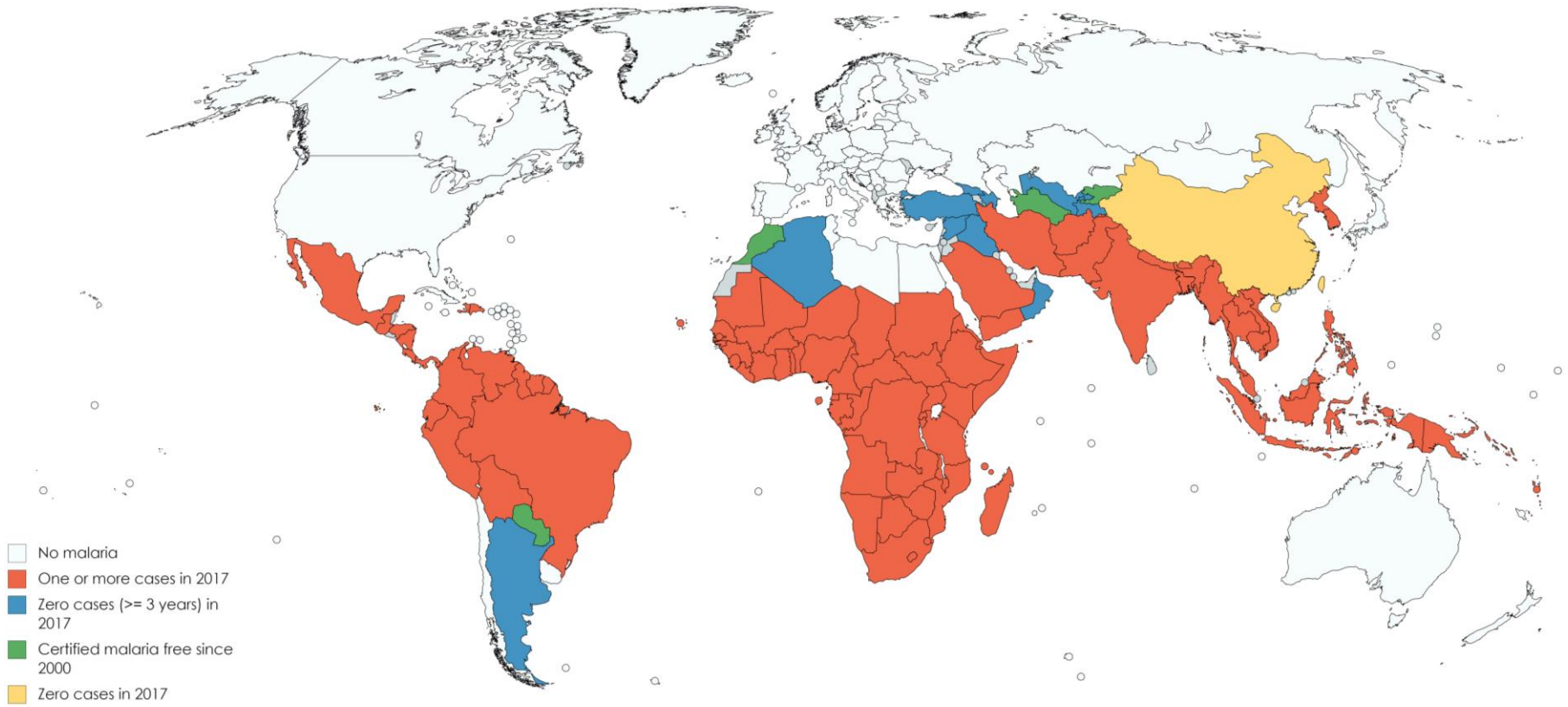


Figure 1.2: Countries with indigenous cases of malaria in 2000 and their status in 2017. Countries that reported no indigenous cases in the past three consecutive years are classified as malaria free. All WHO countries in the European Region report zero indigenous cases in 2016 and 2017. China and El Salvador report no indigenous cases in 2017. Source: World Malaria Report 2018.

1.5 Malaria vaccines

Another major area of research for the prevention of malaria is the development of vaccines. With significant investments in malaria vaccine development research in the past several decades, a recombinant vaccine that incorporates immunogenic epitopes of the circumsporozoite protein, a sporozoite surface protein, called RTS,S vaccine has been developed and tested extensively in clinical trials against *P. falciparum* infection. The vaccine efficacy report for RTS,S is 39%. Although the efficacy of this vaccine is low, it is argued that combining the partial protective effect of this vaccine along with traditional malaria control methods can reduce the adverse effects of malaria[12]. Another vaccine that involves the use of attenuated sporozoites against *P. falciparum* is in clinical trials[12]. In experimental studies, including human volunteer studies, attenuated sporozoites have shown higher efficacy of protection when compared to the RTS,S vaccine. However, the success of this vaccine remains to be tested in future clinical trials.

1.6 Case management methodologies for malaria

1.6.1 Diagnosis of Malaria

Malaria infection status can be diagnosed using various methods. The microscopic examination of Giemsa stained blood smears for the presence of blood-stage parasites is one of the most widely adopted methods for malaria diagnosis[13]. Microscopic diagnostic capacity is limited, especially in Africa. Therefore, immunochromatographic commercial tests that detect one or more of the three parasite antigens, histidine rich protein 2 (HRP2),

aldolase, and parasite lactate dehydrogenase (pLDH), have been widely used for diagnosis[14]. Molecular tests that use PCR based methods are also available, but their use is restricted to research purposes and in reference laboratories for confirmation of diagnosis when required. The WHO policy requires confirmation of malaria diagnosis before treatments can be administered.

1.6.2 Drug treatment

Quinine was one of the first antimalarial drugs used for treatment. Subsequently, at the end of World War II, chloroquine became available for malaria treatment[15]. This was one of the cheapest and most effective drugs used against malaria for many years. As widespread resistance to chloroquine became established globally, sulphadoxine-pyrimethamine (SP) was introduced for primary treatment of malaria. Unfortunately, resistance to SP developed faster than chloroquine, and it became ineffective for the primary treatment of malaria[4]. However, this drug is still widely used for the prevention of malaria in pregnant women and seasonal malaria prevention[15]. When it became clear that widespread resistance to both chloroquine and SP had been established in most endemic countries, the WHO introduced artemisinin combination therapy (ACT) for malaria treatment[15]. This new drug policy was implemented first in Cambodia in 2000 and Peru and Venezuela in 2001. Subsequently, from 2004, other African countries adopted ACT for primary treatment of malaria.

The ACT combines a short-acting drug (half-life 1-2h) and a partner drug with a long half-life. Artemisinin reduces the initial parasite biomass by 95% in 2 days after treatment, and the partner drug eliminates any remaining parasites[4]. Monitoring for the continued efficacy of antimalarial drugs is a key strategy for making sure ACTs are

working as expected in clearing parasites in infected people. There are currently five different ACT combinations, such as artemisinin plus lumefantrine, amodiaquine, piperaquine, mefloquine, SP, and pyronaridine. Although it was hoped that the emergence of resistance to ACT would take longer as two drugs are used, partial resistance to artemisinin was reported as early as 2008 in Cambodia[16]. Recent studies indicate the development of resistance to artemisinin as well as for partner drugs leading to ACT resistance in Cambodia[17]. These developments highlight the importance of continued monitoring of drug resistance in all malaria endemic countries.

1.6.3 Monitoring drug resistance in malaria

The WHO recommends conducting periodic drug therapeutic efficacy studies (TES) in endemic countries, every 2-3 years, to confirm the efficacy of ACTs. The guidelines recommend that drug treatment must lead to the clearance of parasites and no recrudescence (90% or more patients enrolled must remain parasite free) within the observation period (day 28 for most ACTs and day 42 for dihydroartemisinin plus piperaquine). When there is less than 90% efficacy of treatment, resistance is suspected, and further studies must be performed to confirm the resistance. The WHO recommends a change of drugs when resistance is confirmed[18].

In addition to TESs, in vitro drug sensitivity of parasites has been used as a complementary method for confirming resistance. This assay is performed by culturing *P. falciparum* parasites in the presence of a varying concentration of drugs for 2 to 3 days and determines the minimum concentration of drug required to kill 50% of the parasites (IC50). The IC50 of drugs increases several-fold when parasites develop drug resistance. It has

been found that resistance to artemisinin leads to the arrest of *P. falciparum* parasites in the early ring stage and this phenotype is referred to as ring-stage survival assay (RSA) which has been found to correlate with drug resistance[17]. This type of assay requires extensive laboratory capacity in drug trial sites, and it is performed only in some countries[5].

1.6.4 Molecular markers of drug resistance

Development of drug resistance leads to genetic changes in one or more genes that are involved in drug killing pathways (Table 1.1). Identification of these mutations can help identify resistant parasites and can be used as a tool for monitoring the spread of drug resistant parasites.

The *Plasmodium falciparum* Chloroquine-Resistance Transporter gene (*PfCRT*), located on chromosome 7, plays a crucial role in the development of resistance to Chloroquine and Piperaquine. The gene encodes for a drug effluxor protein located on the food vacuole. Chloroquine disrupts the mechanisms through which free haem in the food vacuole is converted to the polymer hemozoin. Mutations in codon 76 (K to T) of the *PfCRT* have been known to confer resistance to chloroquine[15,19–22].

While structurally piperaquine is similar to chloroquine, it has been reported to work against parasites that have *PfCRT* mutations that confer resistance to chloroquine. There have also been studies which report that the presence of the mutations *PfCRT*: C101F in a chloroquine-resistant population, can lead to piperaquine resistance while rendering the parasite susceptible to chloroquine treatment[23–25]. Therefore, different mutations in *PfCRT* regulate resistance to chloroquine and piperaquine.

Table 1.1: Common antimalarial drug and genetic markers associated with drug resistance in *P. falciparum*. *crt*, chloroquine-resistance transporter; *cytb*, cytochrome b; *dhfr*, dihydrofolate reductase; *dhps*, dihydropteroate synthase; *mdr1*, multidrug resistance protein; pfkelch*13*, P. *falciparum* Kelch 13; *plm2*, plasmepsin 2. * Drug used in artemisinin-based combination therapy; Ψ Antimalarial drug used alone or in combination with molecules other than artemisinin derivatives.

Chemical class	Common name	Targeted parasite stage	Genetic marker for drug resistance in *P. falciparum*
Sesquiterpene lactone endoperoxide	Artemisinin*	All parasite stages	*pfkelch13*
	Artesunate*	All parasite stages	*pfkelch13*
	Artemether*	All parasite stages	*pfkelch13*
	Dihydroartemisinin*	All parasite stages	*pfkelch13*
4 - Aminoquinolines	Chloroquine Ψ	Blood stages (trophozoite and schizont)	*pfcrt*
	Amodiaquine*	Blood stages (trophozoite and schizont)	*pfcrt, pfmdr1*
	Piperaquine*	Blood stages (trophozoite and schizont)	*pfplm2, pfcrt*
	Pyronaridine	Blood stages (ring, trophozoite and schizont)	*pfcrt*
	Naphthoquine*	Blood stages (trophozoite and schizont)	Unknown
Amino alcohols	Quinine Ψ	Blood stages (trophozoite and stage I to III gametocytes)	*pfcrt, pfmdr1*
	Mefloquine*	Blood stages (trophozoite and schizont)	*pfcrt*
	Lumefantrine *	Blood stages (trophozoite and schizont)	*pfcrt, pfmdr1*
	Halofantrine Ψ	Blood stages (trophozoite and schizont)	*pfcrt, pfmdr1*
8 – Aminoquinoline	Primaquine*	Blood (gametocyte) and liver (schizont) forms	Unknown
Antifolates	Pyrimethamine*	Blood and liver schizont and mosquito stage (oocysts)	*pfdhfr*
	Sulfadoxine *	Blood and liver schizont	*pfdhps*
	Proguanil*	Blood stages (schizont and gametocyte) and liver schizont	*pfdhfr*
Naphthoquinone	Atovaquone Ψ	Blood stage (schizont and gametocyte) and liver schizont	*pfcytb*
Antibiotics	Clindamycin Ψ	Blood stages	Apicoplast target
	Doxycycline Ψ	Blood stages	Apicoplast target
	Tetracycline Ψ	Blood stages	Apicoplast target

The antimalarial drugs sulfadoxine and pyrimethamine affect the folate pathway in the parasite, by inhibiting two enzymes, *Plasmodium falciparum* dihydropteroate synthase and *Plasmodium falciparum* dihydrofolate reductase, encoded by the gene *PfDHPS* and *PfDHFR* present on chromosome 4 and 8 respectively of the *P. falciparum* genome (Figure 1.3). Mutations in the catalytic sites of these enzymes, and amplification of the two genes have been known to confer resistance to sulfadoxine-pyrimethamine in endemic regions[15,26–30].

The naphthoquinone drug atovaquone, in combination with the proguanil, is administered to people traveling to malaria-endemic countries. The drug combination works by disrupting the mitochondrial membrane potential by targeting the mitochondrial gene *PfCYTb*. Mutations in *PfCYTb* have been found to be associated with resistance to atovaquone[31,32]. Resistance to proguanil has been associated with some mutations in *PfDHFR*[15,31–33].

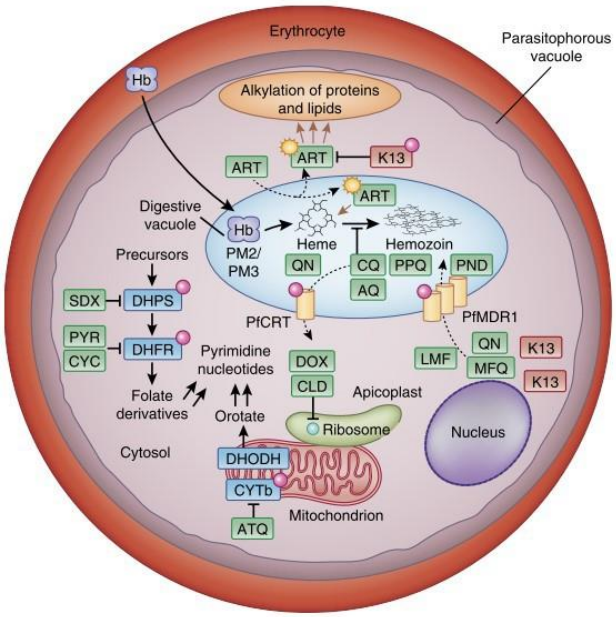


Figure 1.3: The pathways involved in the action of anti-malarial drugs and the molecular markers that affect the resistance. Source: Blasco et al. 2017.

The *Plasmodium falciparum* multidrug resistance (*PfMDR1*) protein is a xenobiotic drug efflux protein that is associated with resistance to many classes of antimalarial drugs (Table 1.1). The gene encoding the food vacuole protein, *PfMDR1*, is present on chromosome 5, and the primary mechanism involved with resistance is through amplification and mutations within the coding region of the gene[5]. Copy number variations associated with *PfMDR1* have been known to confer resistance to many antimalarial drugs including mefloquine, lumefantrine, quinine, and artemisinins[34–37]. Coding mutations in *PfMDR1* such as N86Y, Y184F, C0134S, N1042D, and D1246Y have also been associated with drug resistance[38–40].

Table 1.2: *Candida*te and validated resistance mutations in the *K13* BTB/POZ and propeller domain.

Validated		***Candida*tes/ Associated**	
F446I	P553L	P441L	G538V
N458Y	R561H	G449A	V568G
M476I	C580Y	C469F	P574L
Y493H		A481V	F673I
R539T		P527H	A675V
I543T		N537I	

The main gene associated with resistance to artemisinin compounds is the *PfK13* gene. The gene is located on chromosome 13. This gene is composed of six kelch domains,

and mutations in this region has been found to be associated with delayed parasite clearance[15]. Table 1.2 lists some of the confirmed resistance markers found in *PfK13*. Presence of these mutations at allele frequency > 10% in a given geographical site is considered to be indicative of suspected artemisinin resistance and WHO recommends further investigation to confirm resistance[41]. The exact mechanism by which *PfK13* influences resistances against artemisinins remains elusive.

Molecular markers associated with resistance are useful for tracking drug-resistant parasite populations. Moreover, by studying the genetic variations around drug-resistant markers, one can understand the evolutionary history of drug-resistant parasites. Figure 1.4 shows how resistance to chloroquine (Red arrows), and pyrimethamine (Black arrows) spread globally and describes how ACT resistance is evolving (box) in South East Asia[4]. From microsatellite-based haplotypes alleles flanking resistant genes, it was found that there were only 4 to 5 founding populations of chloroquine-resistant *PfCRT* and they contributed to the global spread of chloroquine resistance. Two such lineages originated in South America, and they spread across the continent. Two lineages from South East Asia contributed to the spread across Asia as well as Africa. Similarly, pyrimethamine resistant *PfDHFR* alleles also spread across the globe.

1.7 Progression from Sanger sequencing to Next Generation Sequencing (NGS) for molecular surveillance

As described in the previous section, molecular characterization of resistance markers is crucial for monitoring resistant parasites as well as for understanding various

aspects of resistance, including evolutionary history and the spread of resistant parasites. Sanger sequencing is the most widely adopted method across public health labs for the characterization of resistant parasites. The cost of analysis using Sanger sequencing, coupled with low throughput and reduced sensitivity at detecting sequence variations at low frequency in mixed infections, make it unsuitable for large-scale surveillance of drug resistance. Advances in NGS methods and the use of targeted genome sequencing approaches has made it very cost-effective to sequence multiple resistance genetic markers in a multiplexed manner, allowing for the identification of minor allele mutations in the population at very low frequency. In this thesis, I develop NGS methods for characterizing six primary *P. falciparum* drug-resistant genes and develop a bioinformatics pipeline to identify well-characterized mutations as well as new mutations in test samples.

1.8 Next Generation Sequencing (NGS) solutions for malaria epidemiology

The improvements in the accuracy of NGS technologies have enabled the use of genomics in better understanding the epidemiology of malaria. In the early days of sequencing, assessments of molecular markers for malaria relied on low-throughput Sanger sequencing-based or array-based protocols to identify specific mutations within samples[42–44]. The decreasing cost of NGS analysis enables large-scale genomic studies to understand linkage disequilibrium, identification of markers involved in pathogenesis, and surveillance of drug resistance[45].

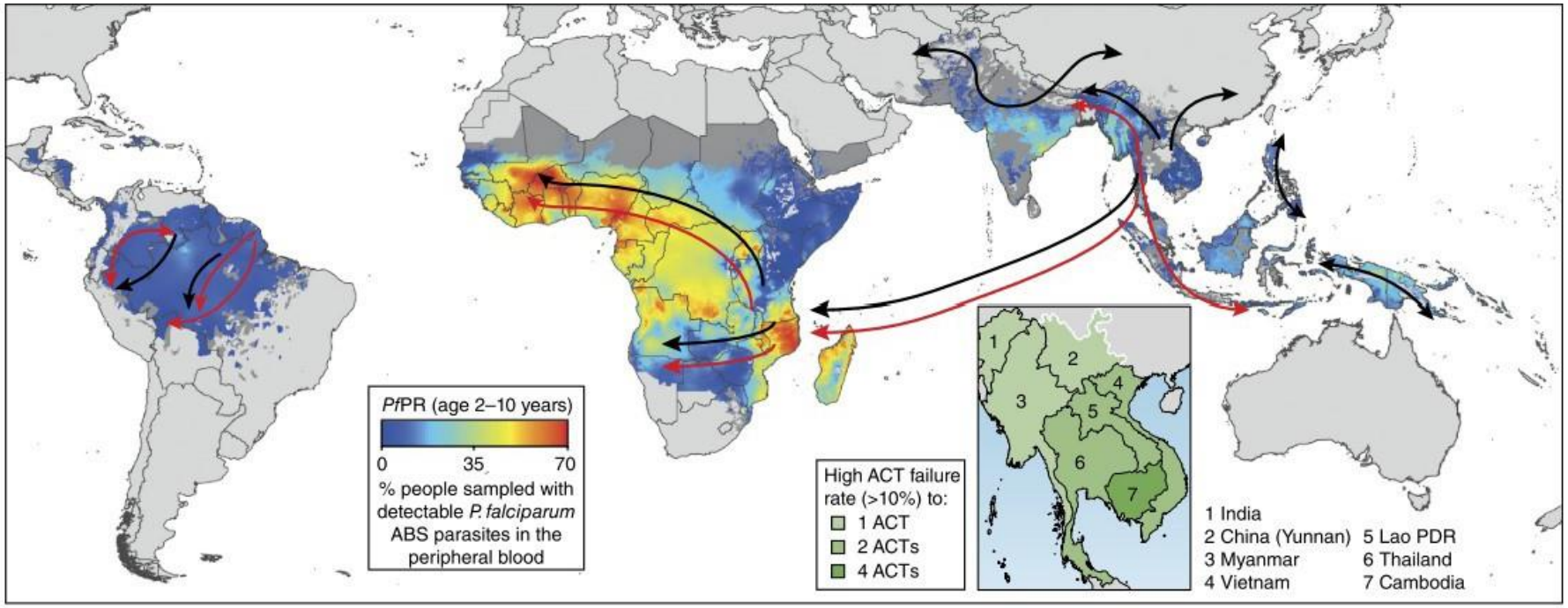


Figure 1.4: Emergence and spread of *P. falciparum* resistance to chloroquine, pyrimethamine, and artemisinin derivatives. Resistance to chloroquine emerged at multiple sites and spread across the world (black arrows), due to the selective pressure on *PfCRT* mutant alleles. Resistance to pyrimethamine emerged in South East Asia and South America. Resistance to pyrimethamine due to a triple mutation in *PfDHFR* spread to Africa (red arrows). Pyrimethamine-resistant *PfDHFR* mutations independently emerged in Africa. Resistance to artemisinin derivatives were driven by mutant *PfK13* alleles and were first detected in South East Asia. Source: *Blasco et al., 2017*

A robust infrastructure needs to be established, one that allows researchers to access and explore these datasets. The malaria research community has been very active at establishing frameworks for malaria data exploration through the concentrated efforts of various consortiums. The Worldwide Antimalarial Resistance Network (WWARN) consortium, established with the support of the Bill and Melinda Gates Foundation in 2009, has developed an online framework to evaluate the efficacy of treatment regimens against malaria, track the prevalence of antimalarial drug resistance, and provide a framework to inform research on the development of new drugs to combat malaria[46].

Another essential resource for malaria genomics has been PlasmoDB[47]. PlasmoDB is a functional database for genomic data, transcript and protein expression data, functional annotation, population genetics, and evolutionary information for *Plasmodium* spp. The information made available to the public through databases like PlasmoDB enable researchers to access to up to date annotations, and reference sequences. Thus, enabling the standardization of large-scale GWAS studies for the evaluation markers associated with the spread of antimalarial drug resistance, as well as the generation of accurate reference genomes through sequence assembly projects.

The MalariaGen consortium provided a significant push towards adopting NGS in malaria epidemiology. With projects such as the Pf3k Project and the Ag100P project, among many others, the consortium aims to utilize Whole Genome Sequencing of Human, Mosquitoes and *Plasmodium* parasites to understand the genetic epidemiology of malaria better. Approaching the same problem from a different angle, investigators in the malaria

branch at the Centers for Disease Control and Prevention (CDC), as a part of the Malaria Resistance Surveillance (MaRS) project, developed an amplicon sequencing protocol, to track molecular markers for drug resistance. The protocol described in Figure 1.5 relies on Target Amplicon Deep Sequencing (TADS) to identify the presence of variants associated with drug resistance across five genes in *P. falciparum* as well as the mitochondrial genome. The goal of the MaRS project is to build a database of variants to study the prevalence of mutations conferring drug resistance, and to identify new molecular markers associated with drug resistance. Chapter 4 of this thesis details the NGS analysis platform, NeST, which is currently being used by the CDC to analyze the data generated from the MaRS protocol[48].

1.9 Algorithms to monitor outbreaks and molecular surveillance of drug resistance using Next Generation Sequencing (NGS).

Building on the established frameworks for epidemiology and molecular surveillance in public health, in this work, I describe two novel algorithms for genomic clustering and molecular surveillance of drug resistance from next generation sequencing data. The two specific aims of the present research are:

1. Development of algorithms for genomic clustering of NGS datasets using an alignment free k-mer based approach.
2. Development of algorithms for the molecular surveillance of drug resistance from NGS datasets in a public health setting.

In the previous sections, the biological mechanisms that govern the spread of drug-resistance in malaria were reviewed. We briefly discussed the molecular markers for drug-resistance, including variants in key genes associated with resistance. Finally, we discussed the tools available for molecular surveillance in malaria.

Chapter 2 reviews existing methods for the identification of Single Nucleotide Polymorphisms (SNP) from sequencing reads and proposes a new consensus-based variant calling framework to overcome the pitfalls of existing methods. State-of-the-art for genomic clustering from NGS data are also discussed here. The basis and advantages of using alignment-free algorithms for genomic clustering is highlighted.

To address the first aim of this thesis, *Chapter 3* outlines the alignment-free k-mer based algorithm, Gentoo, for genomic clustering from NGS data in a public health setting. The new algorithm utilizes k-mer frequencies from isolates to calculate an exact measure of genomic distance. Comparison against the state-of-the-art clustering algorithms highlights the improved resolution of genomic clustering offered by Gentoo. The utility of the method in understanding relatedness of isolates in a public health setting is demonstrated using NGS datasets from *Plasmodium* species and *Candida auris.*

Chapter 4 describes a consensus-based variant calling framework, NeST, to address the second aim of this thesis. The design principles for developing a modular, scalable, standardized framework are described in detail here. The improved precision offered by the consensus framework is evaluated using *in-silico* datasets generated from *Plasmodium falciparum* genes. The applicability of the framework in molecular surveillance of drug resistance is demonstrated using *P. falciparum* samples isolated from imported cases of

malaria in the United States. Finally, the scalability of the frameworks is evaluated by analyzing 8,351 *Mycobacterium tuberculosis* isolates to identify genotypic markers for drug-resistance.

Chapter V concludes this dissertation and summarizes my contributions to computational genomics research in a public health setting.

Figure 1.5: MaRS protocol workflow overview with steps indicated, along with reagents needed, total time, and hands-on time for each procedure.

CHAPTER 2. VARIANT CALLING AND GENOMIC CLUSTERING

2.1 Abstract

DNA sequencing is the process of identifying the order of nucleotides (A, C, G, T) in a molecule of DNA. Knowledge of the sequence of a DNA molecule plays a crucial role in medical diagnosis and epidemiology. The past decade has seen a rapid evolution of DNA sequencing technologies from second-generation high-throughput short-read sequencing to third-generation long-read sequencing methods. The cost of DNA sequencing has been

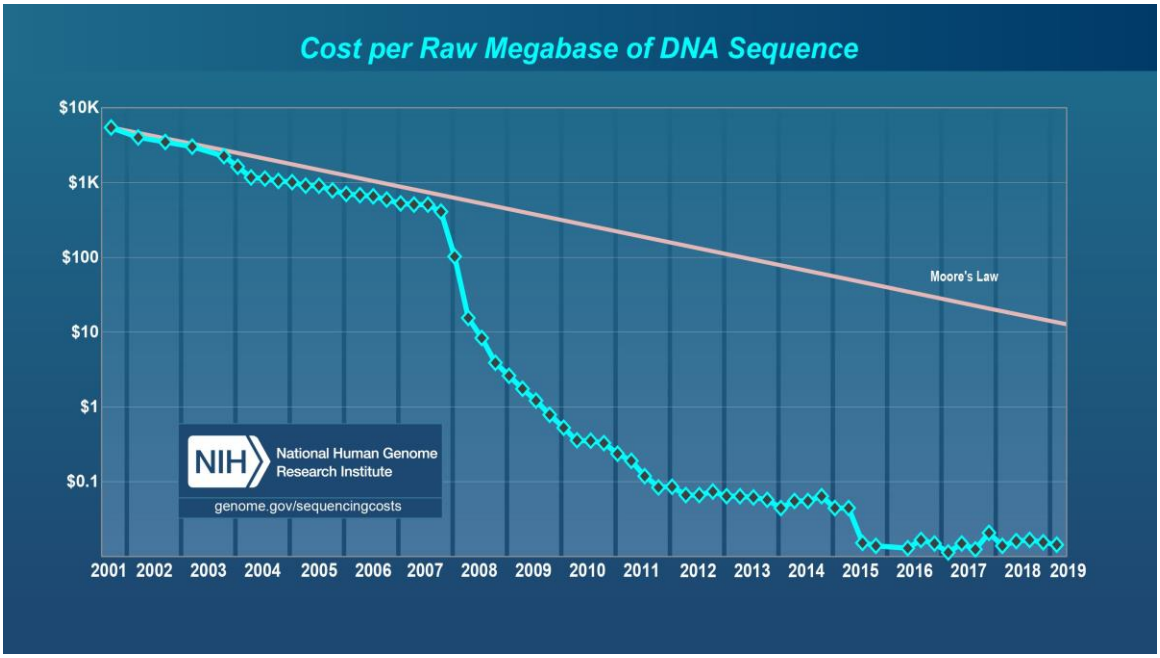


Figure 1.1: Cost per MB of DNA sequenced. To highlight the improvements in sequencing technology, the graph shows a line reflecting Moore's Law, which states that "compute power" doubles every two years. Technologies that keep up with Moore's Law are considered to be doing exceedingly well. The y-axis uses a logarithmic scale. The sudden drop in the cost of sequencing in 2008, corresponds to the switch from Sanger-based to second-generation sequencing technologies.

decreasing at a rate much faster than Moore's Law (Figure 2.1). With such rapid progress, the incorporation of sequencing in biological sciences has been increasing rapidly.

This trend extends to public health as well; an increasing number of studies have been adopting Next Generation Sequencing (NGS) technologies to understand the molecular epidemiology of infectious diseases. A quick search of NCBI's Sequence Read Archive shows that there have been over 70,000 submissions of NGS datasets just for the human malaria parasite *Plasmodium falciparum* to date. With short-read sequencing being the most widely used technology in malaria genomics.

Given the popularity of NGS, the need to develop tools and frameworks that can efficiently and accurately analyze sequencing data is increasing. This chapter outlines two methodologies frequently used in surveillance of infectious diseases, genomic clustering and variant calling from NGS data. The standard protocols and algorithms for each of the methods, their advantages and disadvantages are described in detail. Finally, the underlying principles behind the two new algorithms developed in this thesis and the advances offered by the new methods are discussed here.

2.2 Variant calling from Next Generation Sequencing (NGS) data

Variant calling is the process of identifying Single Nucleotide Polymorphisms (SNPs), and short Insertions and deletions (InDels) in sequencing data based on alignment of sequence reads to the reference genome for that species. Biologists have long relied on sequence variations to understand the factors that are associated with or causative of a disease condition.

As discussed in the previous chapter, the identification of antimalarial drug resistance has long relied on identifying sequence variation. Before the advent of NGS, clinical blood samples containing parasites that displayed antimalarial drug resistance were inoculated into culture to isolate the resistant strain. The resistant strains were then sequenced at the locus associated with resistance using Sanger sequencing techniques to explore the genetic basis of the observed phenotype.

The advent of NGS technologies has enabled the surveillance of antimalarial drug resistance on a much larger scale and with greater accuracy. The ability to multiplex samples on a single sequencing run has drastically reduced the price of sequencing, making NGS based identification of resistance markers more accessible across the world. The main bottleneck is the bioinformatics frameworks available for variant calling.

While there are many methods available for calling variants from sequencing data, most of the methods developed are for model organisms, such as humans and mice. The error correction and filtering of erroneous variant calls rely on large-scale population databases that are not available for most organisms. Due to this, standard tools for filtering low-quality variant calls cannot be used in the context of non-model organisms. Relying on hard filters, based on quality and coverage of sequencing reads for the variant calls, makes it difficult to standardize variant calling pipelines across different studies.

To overcome the drawbacks associated with hard filters, it is proposed here that while the complexity of genomes, the presence of sequencing errors, and algorithmic biases from different methods can lead to erroneous variant calls, true variants in the data should be detected by all variant calling models given high-quality sequence data. Before the

methodology implemented to test this hypothesis is described, it is essential to understand the different steps involved in variant calling as well as understanding the models that are employed to detect active regions and variations using different variant calling methodologies.

The critical steps for pre-processing sequencing reads and variant calling from NGS datasets are described here and three widely used models for the identification of SNPs and InDels are discussed in detail. The methodology implemented in the current research project to overcome the biases of these standard models, is also described.

2.2.1 Steps involved in variant calling

Variant calling from raw sequencing data involves multiple steps to ensure the quality of the data that is being analyzed as well as the accuracy of the alignments being used to detect variants.

2.2.1.1 Quality assessment and control

Base-calling methodologies employed by sequencing platforms rely on signal processing to generate sequencing reads from a DNA template. The accuracy of the base call from sequencing is recorded as a PHRED based quality score and stored in the FASTQ files containing the sequencing reads.

The PHRED score is the negative log-score of the likelihood that the base call is erroneous[49]. The base quality score in the FASTQ files can be used to identify low-quality sequences and filter out sequences that do not meet a necessary threshold for analysis. A commonly used threshold used for quality in short-read data is Q30, or a PHRED score of

30, indicating a 1 in 1000 chance that the base call is erroneous — conferring a 99.9% accuracy for that base call.

Adapter contamination can be another source of sequencing error, which arises when the sequencer reads into the adapter sequence ligated to the DNA fragment that binds to the flow-cell. Adapter contamination usually occurs towards the ends of the sequencing reads.

Quality control tools trim reads with low quality regions and adapter contamination. If reads are trimmed beyond the acceptable length, they are discarded to prevent erroneous alignments or variant calls. If the sequence library is paired-ended, the corresponding paired reads are also discarded. The cleaned reads are ready for the next step of the analysis; alignment against a reference genome.

2.2.1.2 Sequence alignment

A crucial step in variant calling is the alignment of cleaned sequencing reads against a reference genome. The sequence alignment problem is well established, with multiple optimal solutions developed for pairwise sequence alignment. The most popular alignment algorithms are the Needleman-Wunsch[50] for global alignment and the Smith-Watermann algorithm[51] for local alignment. The throughput of NGS data, however, makes it inefficient to use these methodologies for aligning sequencing reads to a reference genome.

Modern sequence aligners rely on seed extension methods to identify the origin of a sequence read and align them accurately, introducing mismatches and gaps when necessary. Most aligners implement an affine gap penalty model that penalizes the

introduction of gaps more than the extension of gap and mismatch penalties to ensure that there is high confidence in any SNPs and InDels identified through sequence alignment. Algorithmic performance is optimized by indexing the reference genome using the Burrow-Wheeler Transform that allows quick access to the reference genome with minimal memory overhead, thus making modern read aligners highly accurate and fast[52,53].

Each aligned read is given a mapping quality score based on the sequence quality, mismatches and gaps introduced, and the number of places the read mapped to, within the reference. A CIGAR string is used to denote the matches, mismatches, and gaps introduced in the sequence. Depending on the aligner used, the user can decide whether local or global alignment of reads is preferred, as well as how many multiple mapping instances are allowed for any given read.

2.2.1.3 PCR de-duplication

One of the steps in library preparation for many sequencing experiments is a PCR amplification step. This can lead to PCR-induced errors in sequence reads where the same fragment is over-represented in the sequence data. This exaggerates the evidence present for a variant or in some cases and causes over-representation of sequencing errors leading to erroneous variant calls. A common approach to resolve this problem is to mark or remove PCR duplicates from the sequence alignment files. Methods that detect PCR duplicates usually identify if the starts and the ends of the sequences are the same. If that is the case, the sequences are marked or removed from the downstream analysis.

2.2.1.4 InDel realignment

Sequence aligners align reads independent of each other. Depending on sequence composition and sequence quality, erroneous gaps can be introduced in low complexity regions. InDel realignment methods overcome this bias by using the alignment of all reads in a given region to assess the placement of a read within the region of interest and realign it based on the evidence found across all alignments in that region.

2.2.2 Variant calling

Variant calling involves the determination of mutations, be it SNPs or short InDels from NGS data from any given sample. Many tools have been developed to perform variant calling on NGS datasets; and they can be broadly classified into three categories.

2.2.2.1 Heuristic models for variant calling

Variant callers implementing heuristic models rely on setting hard thresholds to filter out regions of low-quality or higher noise from NGS data. These thresholds need to be tuned for each dataset, though each tool recommends a default value accounting for the standard error rate from sequencing studies. Following the filtering of low-quality reads, the remaining read evidence is used to identify variants by means of a statistical test, such as a Fisher exact test (used by VarScan2, Shimmer, SOAP-snv), based on the evidence of reference to non-reference bases found in the samples[53–55].

Heuristic models rely heavily on the selection of the right threshold for the filtering of noise from NGS data. When the appropriate thresholds are selected, these methods can be highly accurate at detecting low-frequency variants from NGS data. However, the

reliance on these threshold makes it harder to standardize across sample sets and can lead to erroneous calls when working with low depth datasets[56,57].

2.2.2.2 Probabilistic methods for variant calling

Probabilistic variant calling methods like SAMtools[58], identify the likelihood of each genotype combination based on the sequence evidence present. A likelihood of each genotype is calculated using allele counts, quality scores of each base, and alignment quality at each base in the genome[58]. A posterior probability of each genotype is calculated using the likelihood estimates and an established prior. Either a uniform prior can be used, or a prior can be determined using known population databases such as the dbSNP[59].

The likelihood model is more accurate than heuristic methods when working with low depth samples. The drawback is that probabilistic perform poorly at detecting low-frequency variants. Probabilistic methods also make assumptions of the number of possible genotypes in a given sample, usually assuming a bi-allelic state. The assumption of just two allele states, while simplifying the problem, may lead to a loss of accuracy while making variant calls[60]. Another shortcoming of the position or pileup based probabilistic methods is that they assume independence of bases, leading to erroneous calls when it comes to identifying InDels[56,57,61].

2.2.2.3 Haplotype-based methods for variant calling

Haplotype based variant calling algorithms primarily rely on the same probabilistic framework as the methods mentioned earlier. However, these tools perform a local

assembly around regions of interest to identify SNPs and InDels. Read alignments are used to identify active regions with an increased likelihood of containing a mutation. Reads in and around the active regions are broken into substrings of length "k" or k-mers. These k-mers are used to assemble over the active regions using a De Bruijn graph-like data structure. The assembled haplotypes are then used to determine the likelihood of each genotype within the active regions.

By using local assemblies over alignments, haplotype-based methods such as FreeBayes[60] and GATK HaplotypeCaller[61] can overcome errors in variant calls due to misalignment of reads in low complexity repeat regions in a genome. Since active regions or regions of interest are locally assembled into haplotypes, these methods, in theory, do not need to assume the ploidy or number of copies of DNA for the sample. Haplotype-based callers also enable calling InDels with high accuracy, without the need of the InDel realignment step listed in the previous section[56,57,61].

2.2.3 Consensus-based variant calling

While there are many different methodologies available for variant calling, each method has drawbacks and limitations when it comes to identifying variants from NGS datasets. The degree to which these limitations affect the sensitivity and specificity of the method depends on the quality of the data provided, the genomic complexity of the organism in question, the methodology used for sequencing, sample pre-processing, and downstream filtering of variant calls implemented[62–64].

There has been extensive research done to understand and evaluate the accuracy of the various variant calling pipelines available[62,63,65]. Justin Zook *et al.* conducted a study

in which they sequenced a sample from the 1000 genomes dataset using different sequencing strategies and analyzed the resulting data using multiple variant calling pipelines to identify high-quality consensus variant calls that can be used to validate any given bioinformatic pipeline[66]. Many studies have shown the variance in the calls made by different variant calling algorithms, and conclude that a consensus call using the different algorithms can provide more accurate results[60,61,63,65].

Considering the variance in results from the different algorithms, it is inadvisable to rely on a single variant calling methodology for the identification of mutations from NGS datasets. Moreover, since variant filtering techniques such as Variant Quality Score Recalibrator (VQSR) rely on the availability of population-level databases to identify the accuracy of variant calls, when working with lesser studied organisms such as *P. falciparum*, identification of high-quality variants depends on setting hard thresholds based on sequence abundance, sequence quality, and known error rates of sequencing platforms. Hard thresholds on sequencing characteristics, however, leads to difficulties in standardizing methodologies across different studies since hard thresholds need to be study specific.

An alternative method is to identify high-quality variants using consensus calls from multiple different methodologies. The biases built into each of the different methods can be reduced by considering variant calls that show consensus between different variant calling pipelines. In Chapter 4, a framework that incorporates three different variant calling algorithms is implemented and the accuracy of consensus calls at identifying high-quality variants associated with drug resistance in *P. falciparum* and *Mycobacterium tuberculosis*, is evaluated. Additionally, the design principles that go into building a scalable framework

for consensus calling and the impact it can have in monitoring the spread of drug resistance, are discussed.

2.3 Genomic clustering

The determination of the genomic similarity and transmission history of organisms has been a critical field of research in public health. Originally, the field relied on DNA hybridization techniques to identify similarities between two isolates. However, with the advent of sequencing technologies, new methods were developed to identify relatedness from sequencing data. While evolutionary analysis relies on substitution models to determine the evolutionary distance between organisms, methods for genomic clustering rely on sequence similarity to estimate the pairwise distance between two isolates.

2.3.1 Alignment based genomic clustering

Average nuclear identity (ANI) is the most widely accepted pairwise distance estimation technique for genomic data. Though there have been many implementations proposed to calculate ANI between two organisms[67–69]. The basic concept across all the implementation remains the same. Pairwise sequence alignment is used to identify regions of similarity between a reference and a query given a sequence identity threshold. ANI is calculated as the mean identity of all the similar fragments from the pairwise comparison of the reference and the query[69].

The most commonly used alignment tools to identify pairwise sequence similarity has been BLASTN[70]. Faster algorithms such as Mummer[71], BLAT[72], and DIAMOND[73]

can also be used to determine sequence similarity. ANI has the advantage of being able to estimate relatedness from draft genomes and assembled genomes. However, the reliance on assembled genomes and sequence alignment reduces the scalability of ANI to large-scale genome analysis.

2.3.2 SNP based genomic clustering

A commonly used alternative to ANI for the estimation of the pairwise distance between isolates is SNP based phylogenetic analysis. Here raw NGS data from isolates are aligned to a reference genome, and variant calling is performed on each isolate using methods described earlier in this chapter. Pairwise distance between the organisms is estimated based on the number of SNPs that are shared by the isolates. From the variant calling data, a distance matrix is generated based on the number of variants shared by two samples. This distance matrix can then be used to analyze how these samples cluster, using a neighbor-joining tree[74].

Alignment algorithms for NGS data are much faster as well as much more scalable than pairwise sequence alignment algorithms used for ANI. These algorithms also possess the added benefit of skipping sequence assembly, which can significantly reduce the time taken for the analysis. However, the main drawback of this method is its reliance on a well-established reference sequence, which may not be available in many scenarios.

2.3.3 k-mer based genomic clustering

Many of the drawbacks of alignment-based methods can be mitigated by using alignment-free, k-mer based techniques for estimating pairwise distance. Alignment free algorithms

rely on splitting up the sequencing data in overlapping fragments of the length “k” called k-mers. The overlapping k-mers can then be used to assess the relatedness of samples by considering the extent of shared k-mers between two samples.

Mash[75] and Finch[76], are both alignment-free k-mer based algorithms for the estimation of pairwise distance from raw NGS data. Finch used an XOR Boolean function to measure the pairwise distance from a k-mer occupancy matrix from the two samples. The determination of overlap between samples is done in memory, thus making it a resource-heavy process. Mash, on the other hand, utilizes a MinHash algorithm to calculate a Jaccard similarity score for a given sketch size and k-mer size. The sketch size determines how many k-mers are used to compute the overlap between the organisms. Sampling of k-mers in MinHash based algorithms drastically reduces the memory footprint of the algorithm. Since the distance estimation relies on the sketch size, the Mash distance is an approximate value; an increased sketch size decreases the likelihood of erroneous distance estimation. However, this results in increased memory footprint and run time. Thus, it becomes necessary to find the right compromise for sketch size and resource cost.

Recently Jain *et al.*[77] published a method that uses MinHash to estimate ANI through their implementation of FastANI. The method relies on using MinHash to enable the rapid alignment of sequences as previously implemented in MashMap[78]. The sequence alignments are then used to estimate ANI as per the previously discussed protocol. By speeding up the bottleneck of sequence alignment, FastANI provides a scalable solution for the estimation of ANI from assembled and draft genomes.

In Chapter 3, I introduce a novel alignment-free, k-mer based algorithm, Gentoo, for the estimation of genomic similarity of organisms. The method accepts assembled or draft genomes as well as raw NGS data in FASTQ format. Pairwise distance is calculated using the frequencies or counts of k-mers shared between the sequence datasets of an organism of interest. Having the ability to use raw sequencing data demonstrates its improved utility compared to methods like FastANI[79] and ANI[69]. The utilization of k-mer counts in the estimation of pairwise distance provides a more exact estimate of distance compared to methods like Mash[75] and Finch[76]. The scalability and memory efficiency of the algorithm to accurately cluster isolates is demonstrated using NGS data from *Plasmodium* spp. and *C. auris* samples.

CHAPTER 3. K-MER BASED CLUSTERING ALGORITHMS TO IDENTIFY RELATEDNESS OF SPECIES FROM WHOLE GENOME SEUQENCING DATA

3.1 Abstract

Improvements in short- and long-read sequencing technologies have enabled the use of NGS to determine the relatedness of isolates and species identification in epidemiological settings. Current methods for the determination of relatedness of microorganisms such as ANI[69] and SNP-based clustering methods rely on pairwise alignments of assembled genomes or SNP differences between isolates when compared against a reference genome. While these methods are well-established, they are hard to scale and rely on well-assembled genomes or the availability of high-quality references. While methods like MaSH[75] and FastANI[77] provide faster, scalable alternatives, they rely on probabilistic methods and lose sensitivity when it comes to distinguishing isolates at the species level. Here we propose a k-mer based reference free algorithm, Gentoo, to identify the relatedness of isolates from raw NGS data. In the next few sections, we will describe in detail the algorithm used for the pairwise distance estimation. The accuracy of clustering provided by Gentoo, in comparison with state-of-the-art methods, will be done using NGS datasets from *Plasmodium* spp. genomes. Finally, the utility of Gentoo in the identification of genomic clustering from a real-world outbreak will be evaluated using NGS samples from a *C. auris* outbreak in Colombia. Memory profiling metrics are captured to show the scalability and efficiency of the new method.

3.2 Introduction

In the previous chapter, we discussed in detail the different methodologies available for genomic clustering. The pros and cons of each method were presented, and we touched upon the idea of a reference free, k-mer based scalable tool for genomic clustering and the considerations that go in developing such a system. In this section we detail the implementation of the system and its application in public health. Genomics in public health, especially DNA sequencing, has been mainly used to explore the diversity of infectious species[80,81], explore sequence variations that can be beneficial or harmful to the organism[48,82–84] and use genomic clustering to understand the epidemiological structure of an outbreak[74,85,86].

Improvements in short- and long-read sequencing technologies have enabled the generation of well-annotated complete genome references. Recently, three large-scale genome projects generated complete references for different *Plasmodium* species and strains[80,81,87]. In each case, the knowledge of the closest known species with a complete genome helped with the improvement of genome assembly and annotations produced. While it is easy to establish the closest known relative of well-studied species from their phylogeny, this is a harder proposition for previously unknown or lesser studied organisms. The ability to generate a guide tree from NGS data could help with the identification of the closest known relative of any given isolate and help with the generation of complete genomes.

For evolutionary analysis, phylogenetic techniques usually consider conserved orthologs found across species and apply complex evolutionary substitution models to

arrive at the phylogeny. When dealing with epidemiological studies, looking at outbreaks of infectious diseases, researchers use simpler genomic clustering algorithms to quickly identify the species composition of a group of isolates. The most widely used method for grouping samples by sequence similarity is ANI[69]. The extent of similarity between shared sequences is used to estimate pairwise distance in ANI. Due to its reliance on alignments, ANI is hard to scale to NGS datasets. Other methods such as Mash[75], use probabilistic data structures such as MinHash on sketches of k-mers from the genome, which serve as an approximate representation of the sequence content of an isolate, and estimate distance by calculating the extent of overlap of the k-mer sketches between two organisms. More recently, probabilistic data structures are being used to speed up pairwise alignments, allowing for the use of ANI at a larger scale[77].

SNP-based phylogenetic methods are also commonly used to identify clusters of similar isolates in outbreak scenarios. The distance between samples is estimated by the number of SNPs that are shared by isolates against reference used[74]. While they can provide an effective assessment of the relatedness between isolates, the accuracy of the method is highly dependent on the presence of a complete reference. Here we propose a k-mer based reference-free clustering algorithm, Gentoo, to generate genetic distances from reference genomes, genome assemblies as well as raw FASTQ files from outbreak isolates. The proposed algorithm will calculate pairwise genetic distances from k-mer counts, derived from reference genomes as well as raw NGS data from any sequencing platform. In the following sections, we will highlight the accuracy of Gentoo in building the evolutionary tree of the *Plasmodium* spp. We will evaluate the effect of sequencing errors on clustering accuracy by generating *in-silico* datasets for all the 20 *Plasmodium* genomes. We will

demonstrate the scalability and utility of Gentoo in other outbreak scenarios by clustering isolates from a *C. auris* outbreak in Colombia.

3.3 k-mer based clustering algorithm to identify the relatedness of species from NGS data

Previous work at the Vannberg Lab (Finch) had evaluated the use of k-mer occupancy and Boolean algorithms to estimate the distances between organisms[76]. The previous method accounts for the presence or absence of a k-mer set to evaluate the relatedness. Here we propose a highly parallelized, low memory footprint algorithm that accounts for k-mer frequencies to assess the similarity between two given sets of genomic sequences.

Gentoo accepts reference genome FASTA files, assemblies, and raw FASTQ files. When provided with FASTA files, the k-mers in each contig are counted using KAnalyze[88], by default the k-mer size is set to 31. When working with FASTQ data, the k-mer counting is performed with a filter to remove any k-mer with a count of less than 4. This is set to remove any k-mer generated from sequencing errors.

Following k-mer counting, the files containing the list of k-mers and their counts, k-mer count (KC) files, can be used to calculate the similarity between samples. Given that the count files are numerically sorted, the algorithm scans through two KC files using a merge algorithm. This reduces the comparison problem to an O (N + M) problem, allowing us to maintain a low memory footprint. Gentoo uses the multiprocessing capability of most modern-day computers to spawn multiple threads of pairwise comparisons, thus reducing the overall run-time of analysis.

The metric implemented in Gentoo to evaluate the similarity of organisms is a weighted Jaccard distance (3.1). For each k-mer that is found in both samples, the minimum k-mer count is added towards the intersection set of the organisms. The maximum k-mer count towards the union between the two sets. If a k-mer is only found in one of the two organisms, the count is added towards just the union set.

$$d_{W(x,y)} = 1 - \frac{\sum_i^n \min(x_i, y_i)}{\sum_i^n \max(x_i, y_i)} \tag{3.1}$$

Finch, on the other hand, relied on k-mer occupancy to determine relatedness, i.e., if a k-mer is present in two genomes, it would affect the similarity index for the genomes. Utilizing the k-mer counts as weights will overcome the biases of considering occupancy. It provides a method that is aware of variation in k-mer counts (or coverage at similar regions) and inherently corrects for these biases by assigning a lower weight. In k-mer space variations such as SNPs, InDels, duplications and low-complexity repeats differences between organisms are also represented as differences in k-mer counts of k-mers from the region. Gentoo uses this feature to derive a higher resolution compared to the previously used occupancy method while estimating the similarity between closely related sequences.

All pairwise distances are stored in a distance matrix in memory. Using scikit-bio a neighbor-joining tree is constructed from the distance matrix. The neighbor joining tree is plotted using ete3 toolkit for phylogenetic analysis[89].

Algorithm 1: Calculating weighted Jaccard distance

```
Result: Weighted Jaccard distance between two genomes
intersection = 0;
union = 0;
omer, ocount = next(k-mer, count in Sample1);
tmer, tcount = next(k-mer, count in Sample2);
while (omer != None) and (tmer != None) do
    if omer == tmer then
        intersection += min(ocount, tcount);
        union += max(ocount, tcount);
        omer, ocount = next(k-mer, count in Sample1);
        tmer, tcount = next(k-mer, count in Sample2);
    else if omer <tmer then
        union += ocount;
        omer, ocount = next(k-mer, count in Sample1);
    else if omer tmer then
        union += tcount;
        tmer, tcount = next(k-mer, count in Sample2);
end
while omer != None do
    union += ocount;
    omer, ocount = next(k-mer, count in Sample1);
end
while tmer != None do
    union += tcount;
    tmer, tcount = next(k-mer, count in Sample2);
end
similarity = intersection/union;
distance = 1- similarity;
```

The algorithm mentioned above uses a KC (k-mer count) file as input. As mentioned earlier, in Gentoo, we use KAnalyze[88] to generate k-mer counts from FASTA and FASTQ files. k-mers, as the word indicates, are fragments of a genome sequence of length k. The process of k-mer counting involves breaking a given sequence into fragments using a sliding window across a genomic sequence and counting the occurrence of each k-mer. Since NGS data can have millions of reads, this process can be highly memory intensive and slow. KAnalyze tries to overcome these barriers by using a divide and conquer method. Fundamentally, sequences are first broken into k-mers and stored in segment files of a pre-defined size. Once all the sequences have been broken into k-mers,

the segment files recursively merged and each k-mer is counted along the way. By setting the size of the segment file, one can control the number of k-mers that can be stored in memory at any given point of time and optimize for the runtime of the tool.

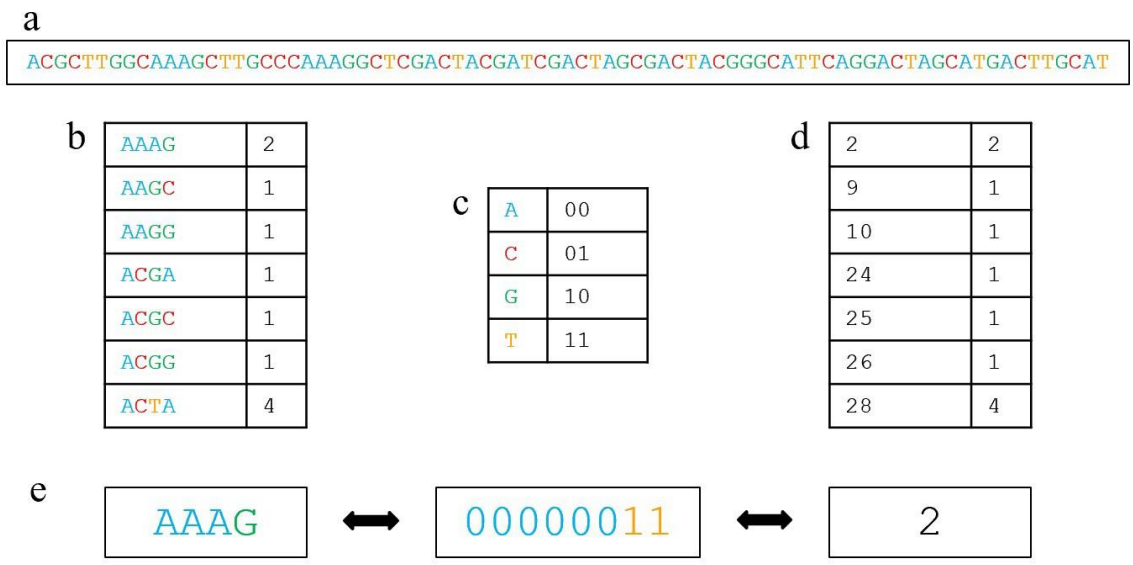


Figure 3.1 Overview of the steps involved in k-mer counting. a) Raw DNA sequences from FASTA or FASTQ files. b) k-mer count tables at 4-mers for the sequence from (a) in string representation. c) 2-bit encoding scheme for DNA sequences. d) Integer representation of k-mer count tables. d) Conversion of string representation of k-mer to integer representation.

KAnalyze also implements an integer hashing function (Figure 3.1) where each base of the DNA is encoded as two-bit integers. Leveraging the fact that with every k-mer we read, only one new base is added, KAnalyze uses a sliding window to scan across a sequence and generate 2-bit encoded k-mers from the sequence data. In most programming languages, integers occupy a fixed amount of memory, the memory footprint of strings, however, varies with the number of characters in the string, the total memory of a string is always the sum of the memory used by each character. Thus, integer encoded strings further reduce the memory footprint of downstream processes. While merging the individual k-mer segment files, KAnalyze implements a merge sort that efficiently counts

and sorts the numerical or lexicographically based on the output format selected. Algorithm 1 can be applied only to sorted datasets, the sorted output from KAnalyze allows for the implementation of a memory-efficient algorithm for downstream processing of k-mer counts.

3.4 Evaluating the accuracy of Gentoo in comparison to ANI as a methodology to identify relatedness of *Plasmodium* spp.

3.4.1 Materials and methods

ANI has long been the accepted standard for the determination of relatedness of organisms from Next Generation Sequencing data, but the reliance on completed assemblies and annotation hinders the scalability of the method. Probabilistic methods like Mash[75] and FastANI[77] on the other hand improve the scalability of the analysis by employing a k-mer based method, MinHash methodology, to speed up the analysis.

To evaluate the accuracy of Gentoo against these methods, 20 genomes from 10 different species of *Plasmodium* studied by Rutledge *et al.*[81] for the phylogenetic analysis of *Plasmodium malariae* were download from PlasmoDB. Nine of the twenty genomes were various strains of *Plasmodium falciparum*.

Further evaluation of Gentoo at estimating pairwise distances from raw NGS data, was done on *in-silico* datasets generated from the *Plasmodium* genomes using DWGSIM[90]. MiSeq paired-end data was simulated for all genomes with error rates of 0.01 and 0.05 (Table 3.1), with a coverage of 30x across the genome. Since Mash and Gentoo were the only tools capable of estimating distances from raw FASTQ files, the trees generated by

Mash and Gentoo from this analysis were compared with the published *Plasmodium* to determine how the methods were able to recreate the ground truth.

ANI values were calculated from the 20 genome FASTA files using PyANI[91], a wrapper around the different steps involved in calculating ANI. Internally, PyANI implements Mummer[71] to identify sequence similarity between the FASTA sequences and calculates ANI from sequences meeting the minimum identity threshold of 90% as described by Goris *et al.*[69].

Mash was run using the default sketch size of 1000 and k-mer size of 21. Mash was run separately on the FASTA and *in-silico* generated FASTQ files. To generate the sketch from FASTQ files, the paired-end sequences for each sample were combined and a sketch was generated using default settings. Pairwise distances between sequences were calculated using Mash dist feature.

Gentoo index was run to generate k-mer counts from FASTA and FASTQ data. K-mer counting on raw FASTQ files was done using KAnalyze[88]. All k-mers of length 31 with counts less than 4 and lowest base quality of less than 20 were discarded. Gentoo cluster was run using the resulting KC (k-mer count) files as input, as described in the previous study, and a neighbor joining tree was generated from the pairwise distances. Since Finch uses the same KC files for clustering, Finch was run on the *Plasmodium* dataset using a k-mer size of 31, as well. A memory profiler was used to record run-time and memory utilization of each of the tools.

Table 3.1: *In-silico* datasets generated from each of the *Plasmodium* spp., genomes.

Sample name	Read count	Error rate
Pbrasilianum	1884594	0.01
PlasmoDB-44_PreichenowiCDC_Genome	1446111	0.01
PlasmoDB-44_Pfalciparum7G8_Genome	1370010	0.01
PlasmoDB-44_PfalciparumGN01_Genome	1422907	0.01
PlasmoDB-44_Pfalciparum3D7_Genome	1400069	0.01
PlasmoDB-44_PvivaxP01_Genome	1744686	0.01
PlasmoDB-44_PmalariaeUG01_Genome	2017480	0.01
Pbrasilianum_draft	1810942	0.01
PlasmoDB-44_PfalciparumHB3_Genome	1368882	0.01
PlasmoDB-44_Pchabaudichabaudi_Genome	1138527	0.01
PlasmoDB-44_PfalciparumCD01_Genome	1412870	0.01
PlasmoDB-44_PfalciparumGB4_Genome	1409521	0.01
PlasmoDB-44_PbergheiANKA_Genome	1126968	0.01
PlasmoDB-44_PknowlesiH_Genome	1464970	0.01
PlasmoDB-44_PfalciparumIT_Genome	1391017	0.01
PlasmoDB-44_PfalciparumGA01_Genome	1388948	0.01
PlasmoDB-44_PfalciparumDd2_Genome	1361026	0.01
PlasmoDB-44_PovalecurtisiGH01_Genome	2013168	0.01
PlasmoDB-44_Pgallinaceum8A_Genome	1503140	0.01
PlasmoDB-44_PcynomolgiB_Genome	1580960	0.01
Pbrasilianum	1884594	0.05
PlasmoDB-44_PreichenowiCDC_Genome	1446111	0.05
PlasmoDB-44_Pfalciparum7G8_Genome	1370010	0.05
PlasmoDB-44_PfalciparumGN01_Genome	1422907	0.05
PlasmoDB-44_Pfalciparum3D7_Genome	1400069	0.05
PlasmoDB-44_PvivaxP01_Genome	1744686	0.05
PlasmoDB-44_PmalariaeUG01_Genome	2017480	0.05
Pbrasilianum_draft	1810942	0.05
PlasmoDB-44_PfalciparumHB3_Genome	1368882	0.05
PlasmoDB-44_Pchabaudichabaudi_Genome	1138527	0.05
PlasmoDB-44_PfalciparumCD01_Genome	1412870	0.05
PlasmoDB-44_PfalciparumGB4_Genome	1409521	0.05
PlasmoDB-44_PbergheiANKA_Genome	1126968	0.05
PlasmoDB-44_PknowlesiH_Genome	1464970	0.05
PlasmoDB-44_PfalciparumIT_Genome	1391017	0.05
PlasmoDB-44_PfalciparumGA01_Genome	1388948	0.05
PlasmoDB-44_PfalciparumDd2_Genome	1361026	0.05
PlasmoDB-44_PovalecurtisiGH01_Genome	2013168	0.05
PlasmoDB-44_Pgallinaceum8A_Genome	1503140	0.05
PlasmoDB-44_PcynomolgiB_Genome	1580960	0.05

3.4.2 Results

Since the completion of the *Plasmodium falciparum 3D7* genome, there have been complete genomes generated for 11 other *Plasmodium* spp. Recently, a large-scale sequencing study used PacBio sequencing to assemble 11 strains of *P. falciparum*[80]. For all these genomes, extensive evolutionary analysis has been performed to determine the phylogeny for the *Plasmodium* spp. as shown in Figure 3.2a. This provides us with a ground truth state to evaluate the different clustering algorithms.

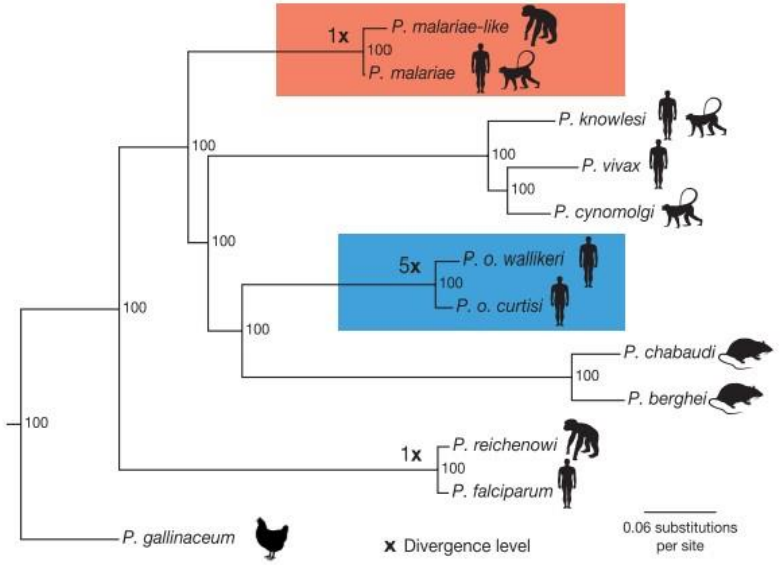
The neighbor joining tree generated from the distance matrix produced by the three methods from genome FASTA files is shown below in Figure 3.2b, 3.2c, and 3.2d. Gen-too, Figure 3.2b, appears to recreate the expected phylogeny from the genome FASTA files, with the exception of the branch point corresponding to *P. ovale curtisi*. In comparison, both trees generated using the distance estimations from ANIm and Mash show erroneous branch points with respect to the expected tree. ANIm places *P. malaria* and *P. brasilianum* in a separate clade as compared to the rest of the species and indicates that *P. knowlesi* and *P. vivax* are closer to *P. falciparum* than to *P. cynomolgi*, contrary to what we observe from the *Plasmodium* spp. evolutionary tree.

Mash on the other hand, correctly places *P. vivax*, *P. cynomolgi*, and *P. knowlesi* as closely related. However, *P. berghei* and *P. chabaudi chabaudi* are grouped with the same clade as *P. falciparum* and *P. reichnowi*, rather than *P. ovale curtisi*, as expected from the *Plasmodium* spp. evolutionary tree.

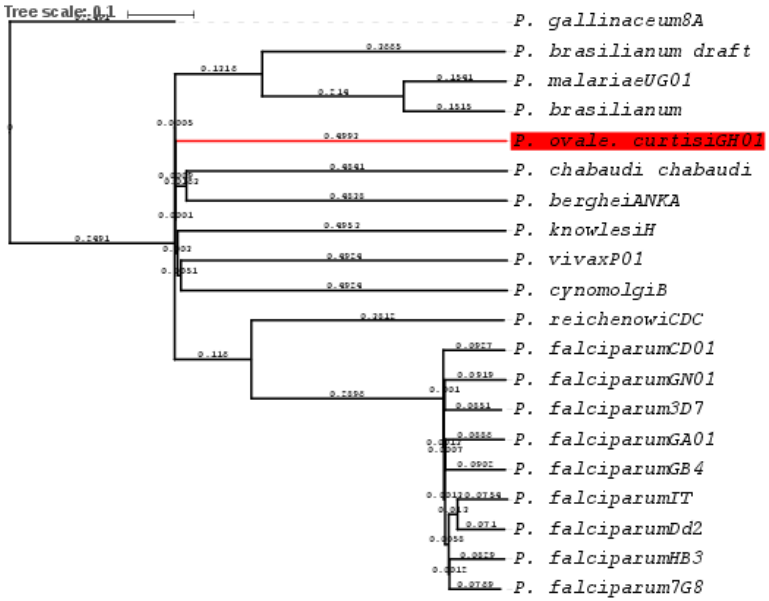
One possible reason for the difference between the expected phylogenetic tree and the trees generated from Mash and ANIm could be the low complexity of the *Plasmodium*

genomes. The genomes from *Plasmodium* spp. are highly AT rich and contain many low complexity repeat regions within the genome. Since there are highly similar repeat regions present across the genomes, pairwise alignments can identify many regions of sequence similarity, but the average nucleotide identity of all the fragments can be low. When it comes to Mash, due to the probabilistic nature of MinHash, the sketch generated for each sample might have a skewed representation of fragments from the genomes due to the increased likelihood encountering a low complexity region in the genome. Since Gentoo uses k-mer counts, overlap of low complexity, high-count k-mers between the weighted Jaccard score, accounts for the abundance of the k-mer, thus normalizing its effect on the estimated distance.

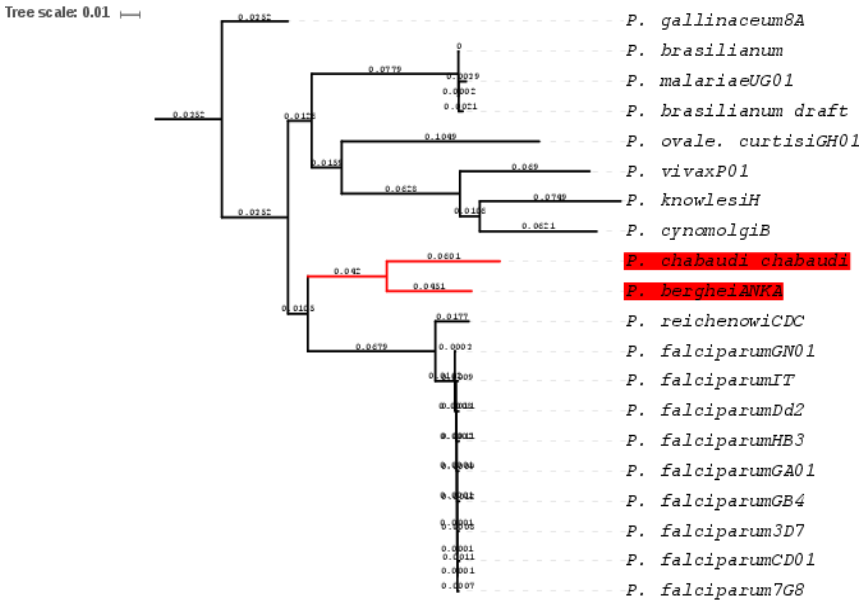
The effect of coverage and error rates on estimation of pairwise distance, was evaluated by running Gentoo and Mash on Illumina MiSeq paired-end data, simulated using DWGSIM[90]. Figures 3.2e and 3.2f show the neighbor joining tree generated from the genomic clustering of the simulated FASTQ datasets from Gentoo and Mash. Here again we observe that Gentoo is able to correctly reproduce the expected phylogeny, with the exception of the branch point corresponding to *P. ovale curtisi*. Since Gentoo utilizes the counts, as well, for the estimation of distances, the resolution of the estimated distances is lower than Mash, but the branch points are mainly as expected for the evolutionary analysis. Mash, however, places *P. berghei* and *P. chabaudi chabaudi* in a separate clade as compared to the other species and places *P. ovale curtisi* closer to *P. brasilianum* than *P. vivax*, contrary to the expected evolutionary tree.



3.2a) Phylogenetic tree for *Plasmodium* genomes.

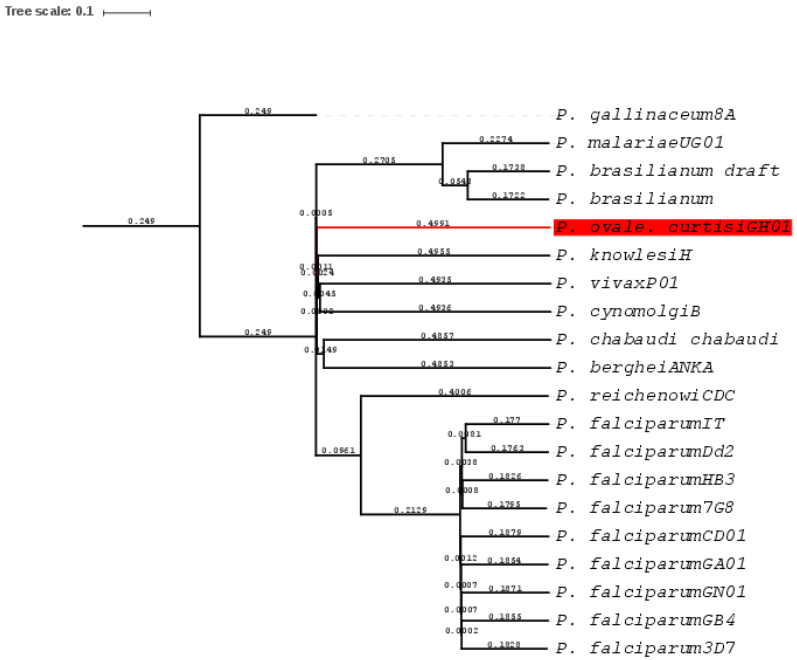


3.2b) Gentoo on 20 *Plasmodium* species FASTA files.

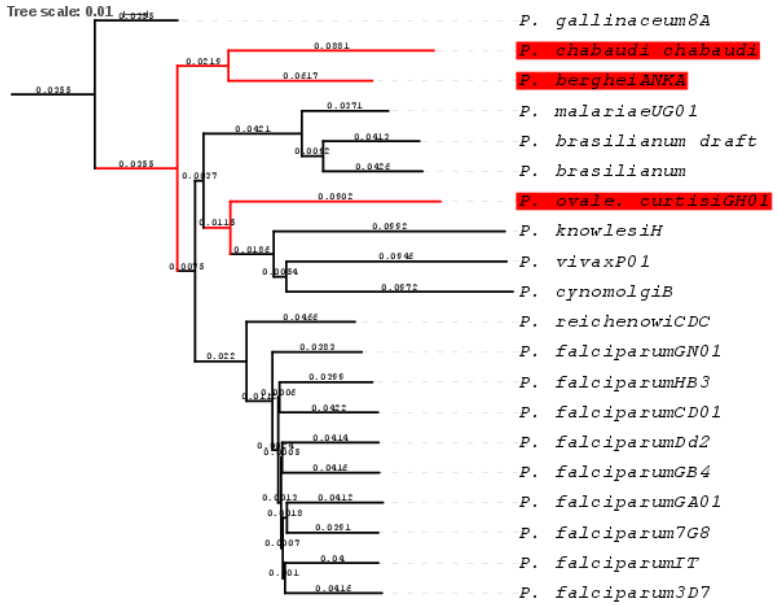


3.2c) Mash on 20 *Plasmodium* species FASTA files.

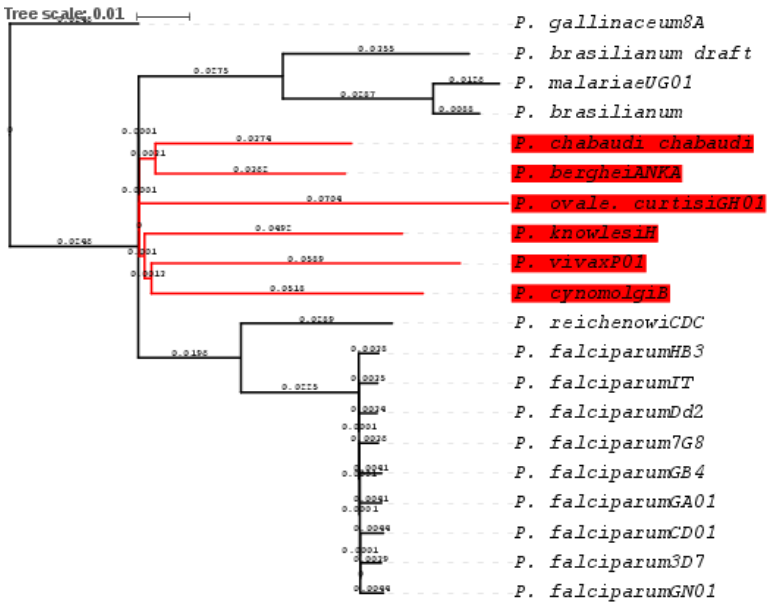
3.2d) ANI on 20 *Plasmodium* species FASTA files.



3.2 e) Gentoo on *in-silico* simulated reads from *Plasmodium* Genomes.



3.2 f) Mash on *in-silico* simulated reads from *Plasmodium* genomes.



3.2 g) Finch on 20 *Plasmodium* species FASTA files.

Figure 3.2: Comparison of neighbor-joining trees generated from pairwise distance estimation made using Gentoo (3.2b, e), Mash (3.2c, f), ANI (3.2d), and Finch (3.2g). Accuracy of the branch points was determined using the *Plasmodium* evolutionary tree (3.2a) published by *Rutledge et al.*,[81]. Branching point comparison shows that the neighbor-joining tree generated by Gentoo is the closest to the *Plasmodium* evolutionary tree, with the exception being the branch point for *P. ovale curtisi*. This hold true even when estimating distances from *in-silico* simulated FASTQ files (3.2e). Tree generated using ANI, Mash and Finch showed at least two branching point errors.

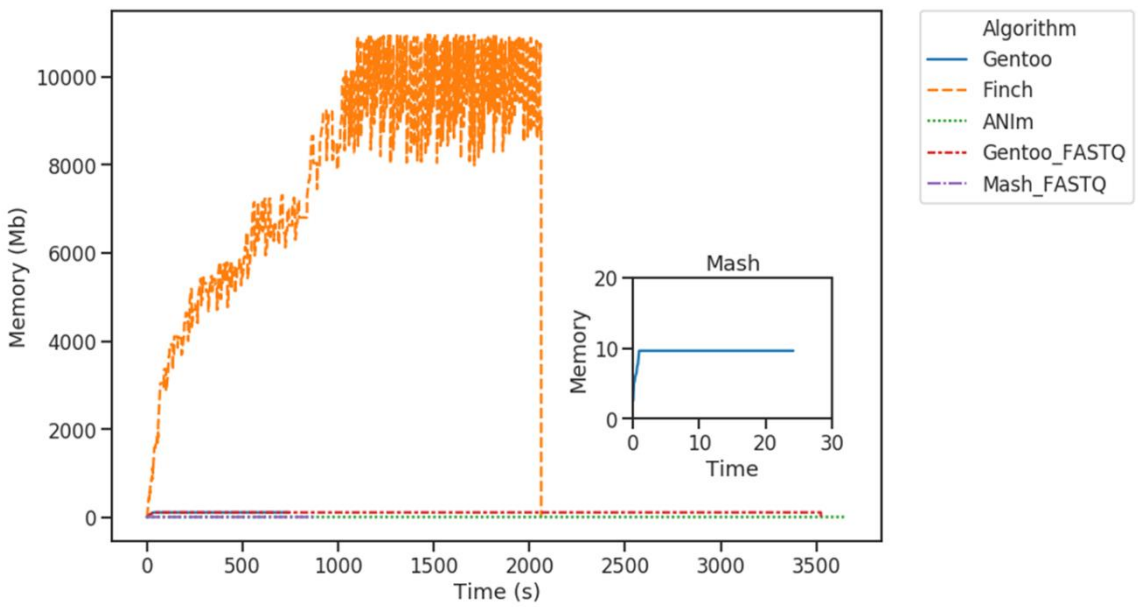


Figure 3.3: Memory utilization by ANIm, Mash, Finch and Gentoo as a functional of time. Memory used by each tool for the clustering of 20 *Plasmodium* genomes was recorded. ANI, Mash and Gentoo were run using 30 concurrent processes. Finch does not allow for the user to specify number of concurrent processes.

The computational resources (i.e., memory, processing time) required for genome clustering were mapped using a memory profiling toolkit (https://pypi.org/project/memory-profiler). The memory usage was plotted as a function of time, as seen in Figure 3.3. As expected, Mash is the fastest and most memory efficient technique when it comes to clustering FASTA and raw FASTQ data. Though the calculation of ANI from assembled genomes has a low memory footprint, the process of generating assemblies from raw FASTQ data for a large number of samples is a memory intensive process.

Although Gentoo is the slowest among the k-mer based methods in this comparison, it has a very low memory footprint. Considering the low memory footprint, the speed of

the analysis can be optimized by adjusting the number of concurrent processes used for the pairwise comparison, in this study we used 30 concurrent processes. Finch on the other hand was the most resource intensive method, as expected, since it loads all the k-mers from each sample into memory for the pairwise comparisons. While this is possible while working with smaller FASTA files, the method is not scalable to larger FASTQ datasets. For these comparisons, all tools were run with 30 threads and 90 GB of RAM[92].

3.5 Clustering outbreaks of *Candida auris* infections in Colombia

3.5.1 Materials and methods

To evaluate the ability of Gentoo to cluster real-world outbreak isolates, we used whole genome sequencing data from *Candida auris* outbreaks in Colombia between 2015-2016[74]. In the study, the authors isolated and sequenced *C. auris* from blood of infected patients. The samples were collected from three hospitals in Bogota, Cartagena and Barranquilla.

The samples were sequenced using the Illumina HiSeq 2500. Illumina reads were aligned to a draft reference previously assembled using PacBio data, and variant calling was performed on the samples. The SNPs were used to construct a maximum parsimony tree. Here, we use these 33 samples to evaluate the ability of Gentoo to recapitulate the epidemiological data from[74]. FASTQ files were indexed using KAnalyze[93] with a k-mer size of 31. A neighbor joining tree was generated from the indexed FASTQ files.

3.5.2 Results

We downloaded 33 *C. auris* isolate that were sequenced from blood samples of patients at three hospitals in Bogota, Cartagena, and Barranquilla during a suspected *Candida auris* outbreak[74]. The raw data for these samples was downloaded from the NCBI Bioproject PRJNA470683.

K-mer counting on the raw FASTQ files was done using KAnalyze[88]. All k-mers of length 31 with count less than 5 and lowest base quality of less than 20 were discarded. Gentoo was run using the resulting KC (k-mer count) files as input. Pairwise comparisons between all 33 samples were performed using the algorithm described in Algorithm 1. A distance matrix was created from the weighted Jaccard scores, and a neighbor-joining phylogenetic tree was constructed for the outbreak isolates (Figure 3.3a). The results from Gentoo were compared with the published phylogenetic tree from[74] (Figure 3.3a). In Figure 3.3a we see that isolates from hospital A cluster together; however, isolates from hospitals B and D form two distinct clades.

The phylogenetic tree generated using Gentoo (Figure 3.3b), however, groups almost all the samples specific to their corresponding geographic origin, with the exception of samples B1156D (hospital D) and B11846A (hospital A). Though Gentoo is able to group isolates from each hospital correctly, the placement of samples from hospital A with samples from hospital D is at odds with the fact that hospital A is in northern Colombia and hospital D is in central Colombia. This discrepancy could be due to the noise arising from using k-mer profiles of the raw FASTQ files. Optimizing for k-mer size and minimum base quality of k-mers used for the analysis might resolve this discrepancy. The major

advantage of using an alignment free method, such as Gentoo, over a SNP-based distance estimation technique, is the significant gain in speed. Moreover, an alignment free approach allows for clustering sequences without the need of a reference genome.

3.6 Discussion

Estimated distances calculated between any two genomes may be skewed by depth of coverage when using raw NGS data versus assembled genomes. Here we demonstrate that using k-mer counts while clustering NGS datasets can overcome the effects of depth and sequence complexity in the calculation of pairwise distances. We demonstrated this by comparing Gentoo with other state-of-the-art methods for genomic clustering, such as ANI, Mash, and Finch to recreate the evolutionary tree for *Plasmodium* spp.

Considering the AT-rich nature of the genome and high frequency of repeat sequences, percent identity-based methods, such as ANI, and k-mer occupancy-based methods, such as Finch and Mash, can produce results inconsistent with the expected phylogeny. The use of count information in the calculation of pairwise distance, as implemented by Gentoo, genomic difference and repeat composition between species to provide a better estimation of the phylogeny. We further showed that the count-based distance allows for accurate clustering from raw FASTQ data, without any prior error correction or assembly. This can prove extremely useful since genome assembly of large genomes is still a time and resource intensive process.

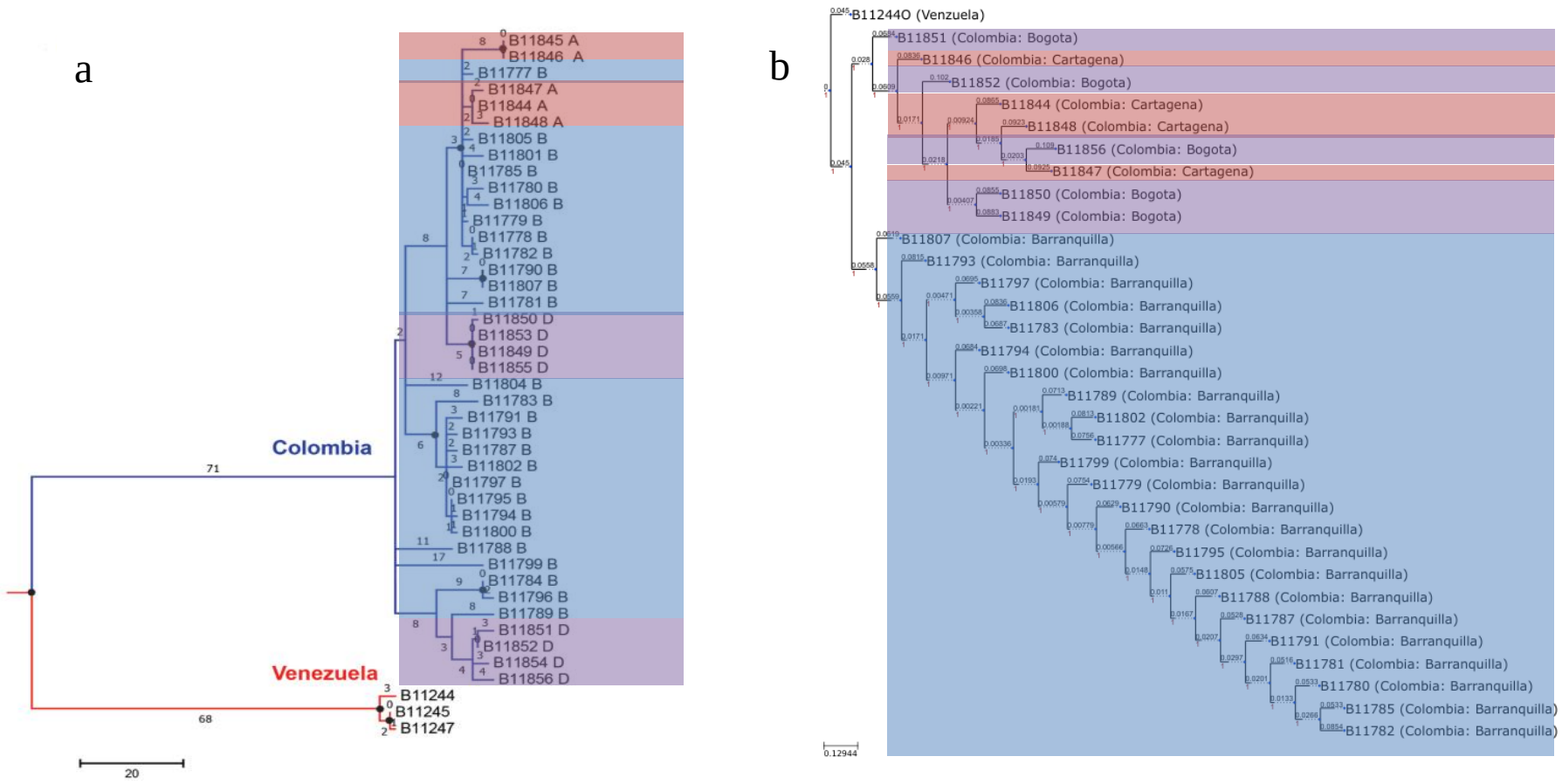


Figure 3.4: Tree generated from *C. auris* outbreak isolates from Colombian and Venezuelan isolates (BioProject ID: PRJNA470683). Clades are shaded based on the geographical locations from which the samples within the clade were isolated. Samples shaded in red were isolated from Hospital A in Cartagena, those shaded in blue were isolated from Hospital B in Barranquilla, and those in purple were isolated from Hospital D in Bogota. a) Maximum parsimony tree of *C. auris* isolates. b) Neighbor-joining tree generated using Gentoo.

The applicability of using Gentoo in a real-world outbreak setting was evaluated by clustering *C. auris* isolates from an outbreak in northern and central Colombia. While Gentoo was able to cluster isolates from the same hospital more accurately in comparison to SNP based phylogeny, both SNP based phylogeny and Gentoo failed to cluster the samples by their geographic distribution.

While this can be a comment towards the accuracy of the method, real world samples have a level of noise associated with each sample in terms of parasitemia (i.e., the admixture within the sample), and clustering algorithms may not be able to provide clear resolution at a geographic scale. Sequencing data from isolates cultured prior to sequencing can provide greater resolution as we see with *C. auris* isolates; however, complete geographic resolution is very difficult to achieve.

Resolving the admixture in samples before clustering might achieve a better resolution. While there have been methods proposed for de-convolution of infection isolates from NGS data[94,95], there is a long way to go before NGS data can be used to achieve perfect resolution of genomic clustering from blood isolates in outbreak scenarios. Additionally, when considering isolates from cases involving inadvertent contamination of biologics or in the case of bioterrorism, isolates are much more likely to show a greater level of clustering due to the clonal nature of the population, demonstrating the potential use of genomic clustering in outbreak and bioterrorism scenarios.

CHAPTER 4. NEXT GENERATION SEQUENCING AND BIOINFORMATICS PROTOCOL FOR MALARIA DRUG RESISTANCE MARKER SURVEILLANCE

4.1 Abstract

Recent advancements in Next Generation Sequencing (NGS) and bioinformatics along with decreasing costs per base sequenced have led to wider adoption of these methods in public health settings. While there is an abundance of protocols for NGS and bioinformatics analysis, many are tailored for research purposes and model organisms with extensive population-level data. These protocols are now being evaluated, modified, and standardized for routine use in public health laboratories. The vast majority of public health laboratories utilize the Illumina short-read sequencing technology and various different algorithms or variant callers for Single Nucleotide Polymorphisms (SNPs) based data analysis. In an effort to standardize SNP based analysis and overcome the inherent biases of any individual SNP based variant caller, a Next-generation Sequence analysis Toolkit (NeST) was developed. NeST provides a modular consensus-based variant calling framework for the identification of SNPs and short Insertions and Deletions (InDels). NeST uses a combination of variant callers that provide metrics to assess the accuracy of a variant found in a sample. NeST consists of four distinct modules: (1) PrepInputs, (2) VarCallEngine, (3) VCFToolkit, (4) Summarize. The utility and scalability of NeST is demonstrated by its recent adoption at the CDC for the molecular surveillance of malaria parasites[48]. In addition, using in silico data sets and *Mycobacterium tuberculosis* whole genome sequencing (WGS) data, we assess NeST's accuracy, sensitivity, and specificity as compared to other SNP based algorithms.

4.2 Introduction

The spread of antimalarial drug resistance is a major threat to controlling and eradicating malaria. As described in Chapter 1, antimalarial drug resistance can be linked to mutations in crucial genetic markers in *Plasmodium falciparum*. Molecular surveillance of drug resistance relies on tracking the prevalence of these mutations in a given population.

Next Generation Sequencing (NGS) has had an enormous impact on molecular surveillance of drug resistance. With the improvement in second-generation NGS technologies, including throughput and reduced cost of sequencing, routine NGS based molecular surveillance of drug resistance in malaria is becoming more widely adopted. Second-generation Illumina short-read NGS is now one of the most widely used sequencing techniques in both research and public health sector. Short-read sequencers generate a large amount of high-quality sequencing data and can provide a cost-effective method for the surveillance of drug resistance for hundreds of samples by taking advantage of the ability to multiplex samples in a single run. The major bottleneck, however, is the availability of standardized, consensus-based bioinformatics tools to analyze this data.

Genotypic determination of drug resistance mainly relies on the identification of Single Nucleotide Polymorphisms (SNPs) and short Insertions and Deletions (InDels) in genes identified as key markers for drug resistance[15,84,85]. Accuracy of the predicted genotypic markers for resistance is contingent on the accuracy of the variant calls made from NGS data.

A large number of tools and pipelines have been developed for variant calling from NGS data[60,64,65,96–98]. Studies have shown that there is a considerable amount of

discrepancy in the variant calls made from different algorithms[63], leading to is an increasing reliance on variant filtration algorithms to select high-quality variants. Variant filtration algorithms utilize true variant calls from the population and apply machine learning to identify high-quality calls from the sample data[98]. For most organisms, however, population-level information is not available.

Hard filters on variant calls have been suggested to circumvent the lack of population-level information, but they are hard to standardize across sequencing protocols[84]. Having a consensus variant call that relies on different algorithms can provide an alternate discrete metric to determine the quality of the calls made.

In this chapter, we introduced a novel Next Generation Sequencing analysis toolkit (NeST) for the identification of high-quality consensus calls from NGS datasets. First, we will describe in detail four key modules that make up the framework and detail the standardization of inputs and results generated by NeST.

Next, we will demonstrate the advantages of using a consensus framework over individual pipelines by studying the variability of the results produced by different variant calling methods using *in-silico* datasets. Third, we will describe the utility of the framework by highlight the implementation of NeST as a framework for the surveillance of antimalarial drug resistance at the Centers for Disease Control and Prevention (CDC). Finally, will show the scalability and adaptability of the framework to other outbreak scenarios by evaluating its accuracy at identifying mutations conferring drug resistance in *Mycobacterium tuberculosis*.

4.3 NeST variant calling framework

Next-generation Sequence analysis Toolkit (NeST) (https://github.com/shashidhar22/NeST), is a modular consensus-based SNP calling framework for variant calling that integrates open-source bioinformatics tools for quality correction, alignment, and SNP calling using NGS data. The key design consideration made during the development are i) Reproducibility, ii) Cross platform compatibility, iii) Usability, iv) Modularity, and v) Scalability. These principles are implemented through four key modules in NeST.

4.3.1 PrepInputs Module

To improve reproducibility and usability, for each study NeST generates a study object containing all the information regarding the samples, the reference genome, gene boundaries, and variants of interest for the study provided by the user. The sample data can be provided to NeST in three formats: a path to the folder containing the raw FASTQ files, an SRA accession list or a tab-delimited list with sample names and associated FASTQ files or SRA accession number. The module downloads the FASTQ files using SRAToolkit[99], merges technical replicates and collects all the files required for the processing of samples in a study. The module parses each FASTQ file associated with a sample and uses the FASTQ headers to retrieve the relevant sequencing run information, including sample name, library type, and sequence length. FASTQ-files are grouped by sample name in a run dictionary before initiation of the analysis.

4.3.2 VarCallEngine Module

Three sets of SNP calls are generated for each sample by the VarCallEngine module. Sequencing reads are first trimmed and cleaned based on pre-defined quality thresholds, and adapters removed using BBDuk[100]. Cleaned reads are then aligned to a reference genome using one of the four included aligners (BWA[53], Bowtie2[52], BBMap[100], SNAP[101]). If none are specified, BWA is run using default settings. The alignments are then sorted, de-duplicated, and read group information added using SAMTools[58] and Picard (https://broadinstitute.github.io/picard). The de-duplicated BAM files are then used for SNP (Single Nucleotide Polymorphism) calling using Freebayes[60], GATK[61] and SAMtools-BCFtools[58]. Due to variation in the representation of InDel by different tools, as of the current version, NeST performs a consensus variant calling only on SNPs. InDels are however reported as they are found in each of the callers.

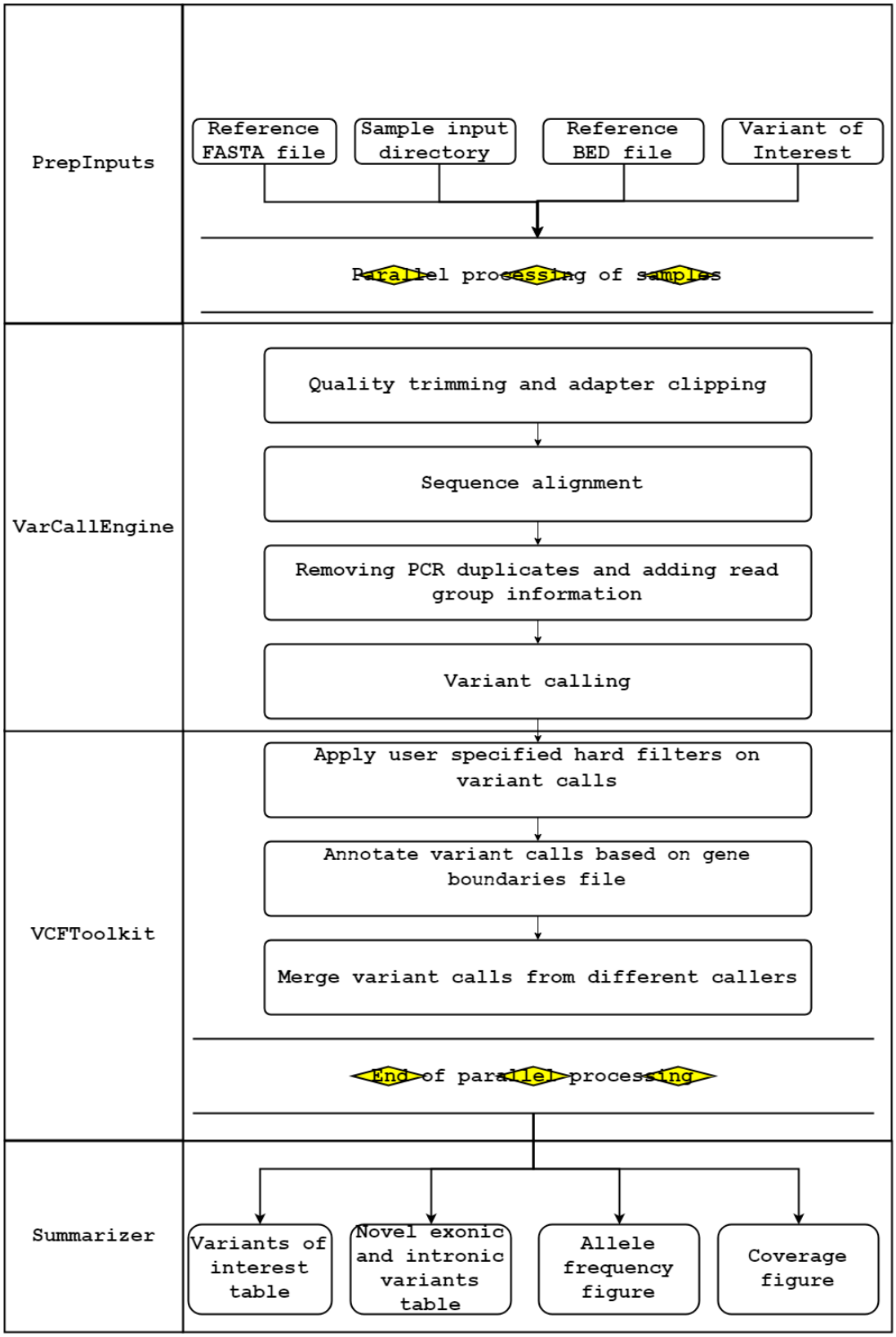


Figure 4.1: NeST flowchart detailing the four main modules. PrepInputs consolidates all user inputs. VarCallEngine executes three variant calling pipelines. VCFToolkit annotates the variant calls and merges VCF files to provide a consensus variant call. Summarizer generates human readable reports and figures from the NGS analysis.

4.3.3 VCFToolkit Module

The VCFToolkit Module is a custom Variant Call Format (VCF) file parser used to filter, merge, and annotate variant calls from the different variant calling algorithms implemented in NeST. The parser is composed of sub-modules, allowing for further customization.

The filter module allows the user to filter SNPs by standard VCF fields. The VCF files are then annotated using a BED file, with designated gene boundaries, provided by the user. BED files are easy to modify and allow for easy annotation of VCF files for organisms that lack an annotated gene or genome database.

The variant calls are then combined using the merge module. This module sequentially parses through the VCF files and merges all headers, INFO, and FORMAT field values in the VCF file. The INFO field called Confidence is added, which indicates the number of variant callers that identified a SNP. An additional Sources field is added to indicate the callers that identified each variant. A list of VCF files are provided to the merge module and are split into pairs and recursively merged. Thus, allowing for easy merging of results from multiple variant callers.

4.3.4 Summarize Module

The summarize module combines and compares the annotated SNPs for each sample from the VCF files. A data-frame is created showing the presence of known SNPs (i.e., user-defined via the reportable SNPs document) across all samples. Non-user defined novel exonic and intronic SNPs are grouped into two separate tables (Table 4.4). Custom R

scripts then automatically generate figures summarizing the sequencing depth (Figure 4.3) and allele frequency (Figure 4.4).

4.3.5 Accessibility and cross platform compatibility

As we have seen in the previous section, NeST stitches together various open-source tools, threading the results from one tool to the next, and automating the whole process. The increased number of dependencies creates multiple failure points within the framework. It is essential to keep this in mind while designing any bioinformatics framework, that most of the tools used are open source in nature. The dependencies usually also vary in the language used to develop the tool, the dependencies used by the tools, as well as the frequency with which the tools are updated or maintained. Table 4.1 lists the various design considerations that went into the evolution of the NeST variant calling framework. In this section, we will focus on virtual environments and cross-platform compatibility of the framework.

NeST relies on the Anaconda virtual environment to ensure version control of the tools used within. The Anaconda installation provides a framework to maintain exact versions of dependencies required for any analysis. Installation of most tools requires either admin or superuser privileges depending on the platform being used. Virtual environments, as the name suggests, create a virtual space within the users' profile. Here the user can install, modify, and delete any package without affecting the system environment that is shared by all the users. This allows the user to maintain many different versions of dependencies and software stacks without needing admin privileges.

Table 4.1: Comparison of usability features available across open-source variant calling platforms.

Feature	NeST v2	NeST v1/MaRS	CoVaCS	Omics Pipe	NARWHAL
Consensus variant calling	✓	✓	✓		
Open source	✓	✓		✓	✓
Multi-sample analysis	✓	✓	✓	✓	✓
Automated FASTQ retrieval from SRA	✓	✓	✓		
Amplicon Sequencing Data	✓	✓		✓	✓
Automated figure generation	✓	✓			
Summarization reports	✓	✓			
Cross platform compatible (Linux, OSX, WSL)	✓	✓		✓	✓
Local installation	✓	✓		✓	✓
Cloud deployable	✓	✓	✓		
Whole Genome Sequencing Data	✓		✓		✓
Pathogen agnostic	✓			✓	✓
Modular variant annotation toolkit	✓				
HPC deployable	✓			✓	
Automated installation	✓			✓	
Virtual environment with version-controlled dependencies	✓				

```yaml
channels:
  - conda-forge
  - defaults
  - bioconda
  - r

dependencies:
  - fastqc=0.11.8
  - perl=5.26.2
  - r=3.5.1
  - igv=2.4.6
  - python=3.6
  - pysam=0.15.1
  - pandas=0.23.4
  - numpy=1.15.2
  - xlrd=1.1.0
  - openpyxl=2.5.8
  - bbmap=38.22
  - bwa=0.7.17
  - bowtie2=2.3.4.3
  - samtools=1.9
  - bcftools=1.9
  - bamtools=2.5.1
  - gatk4=4.0.4.0
  - picard=2.18.14
  - sra-tools=2.9.1_1
  - readline=7.0
  - r-optparse=1.6.1
  - r-ggplot2=3.0.0
  - r-dplyr=0.7.6
  - r-tidyr=0.8.1
  - r-readr=1.1.1
  - r-stringr=1.3.1
  - r-RColorBrewer=1.1_2
  - r-getopt=1.20.2
  - jupyter
  - scikit-learn
```

Figure 2.2: NeST virtual environment.

The virtual environment is usually controlled using a YAML file. Figure 4.2 shows the YAML file used to generate the NeST virtual environment. The channels mentioned in the YAML file maintain the version of the dependencies listed. Miniconda downloads the required version of each of the dependencies and installs them to the local virtual environment. The bioinformatics tools used in NeST are maintained by BioConda[102]. The advantage of using a virtual environment is the cross-platform compatibility that is offered. Miniconda can be deployed on Linux and OSX frameworks, as well as The Linux subsystem for Windows. Thus, allowing users on all platform access to the framework.

4.3.6 Input and Result standardization

In this section, we will describe the various inputs that NeST requires for any analysis as well as describe the outputs generated from the analysis. NeST is designed to reduce the amount of user intervention with regards to inputs that the user needs to provide. However, to enable standardization of inputs across all organisms, we require that a particular file format be followed for the three inputs listed below:

1. FASTQ files:

The PrepInputs module in NeST highly simplifies the management of FASTQ files. The module accepts two input formats.

   a. Input directory path:

The user provides the path to a folder containing FASTQ files. The files with the path are recognized by the file extension. The allowed extensions include fq, fq.gz, fastq or fastq.gz. The naming convention followed for paired sequencing read files must be _1, _r1 or _R1.

b. SRA accession list:

The user also has the option to provide a text file with a list of SRA experiments, with one SRA number per line of the file. This can be exported from the SRA run selector tool. An example SRA accession file is provided with the NeST installation.

2. BED format:

The BED (Browser Extensible Data) is an easy and lightweight format to list annotations for a genome. NeST uses a full BED or BED 12 column format file as a guide to annotate variants with codon and amino acid changes. The separation of contig, gene and exon level information make this format highly portable across genomes. The BED 12 column format for most organisms can be export from the UCSC table browser.

3. Variants of Interest:

The Summarize module in NeST allows for collates all the variants (SNPs and InDels) called from all the samples in the study. If a user specifies a list of variants of interest, a separate table will be created for these variants. The variants can be in comma separated, tab separated or excel format. Table 4.2 provides an example of the variant of interest file.

Table 4.2: Variants of interest table. Each row should contain the Chromosome, Gene name, reference amino acid, alternate amino acid and the amino acid location for the variant of interest.

Chrom	Gene	RefAA	AAPos	AltAA
PfCRT	PfCRT	C	72	S
PfCRT	PfCRT	V	73	V
PfMDR1	PfMDR1	N	86	Y
PfMDR1	PfMDR1	Y	184	F
MT	CYTOb	I	258	M

Each NeST analysis produces a standard list of tables and figures that summarize the data from the variant calling experiment. Below we describe the different output that are generated by NeST.

1. Report files:

NeST generates tables that summarize the different types of variants found in the samples. All the tables will be stored under the Reports folder inside the output directory. Table 4.3 describes the different files that are generated by NeST.

Table 4.3: List of summary files generated by NeST.

File	Description
Study known variants	This file contains the calls for each of the variants of interest, for each of the samples. The table also lists the variant call metrics for the variants
Study known variants allele frequency	This file lists the allele frequency for each of the variants of interest, for each of the samples in the study
Study known variants depth	This file lists the depth for each of the variants of interest, for each of the samples in the study
Study novel exonic variants	This file lists all the novel exonic variants found in all the samples in the study along with the variant call metrics
Study novel intronic variants	This file lists all intronic variants found in all the samples in the study along with the variant call metrics
Study novel variants allele frequency	This file lists the allele frequency for each of the novel variants, for each of the samples in the study
Study novel variants depth	This file lists the depth for each of the novel variants, for each of the samples in the study

2. Figures:

a. Study Depth:

Read depth of coverage for SNPs associated with drug resistance. SNP loci are shown on the on the x- axis, and the read depth of coverage on the y-axis. The colors indicate the genes that were amplified during sequencing (Figure 4.3).

b. Reportable SNPs:

Bar graph depicting the wild type, major and minor allele frequencies of associated and/or confirmed resistance SNPs. Allele frequencies are indicated on the x-axis, and the variants of interest are listed along the y-axis (left). The number of samples that had a particular mutation is indicated on the y-axis (right). The color coding indicates the type of mutation found in the samples; blue is for wild type, green for minor allele mutation and red for major allele mutation (Figure 4.4).

c. Novel Intronic SNPs:

The figure follows the same format as Figure 4.4. The mutations indicated on the y-axis are any mutation that did not lie within the gene boundaries defined in the BED file, i.e., intronic and intergenic mutations.

d. Novel Synonymous Exonic SNPs:

Novel synonymous exonic SNPs are also reported in the same format as Figure 4.4. The mutations indicated on the y-axis are synonymous SNPs that lie within the exons of the genes of interest and have not been specified in the variants of interest file, provided by the user.

e. Novel Non-Synonymous Exonic SNPs:

Following a similar format to Figure 4.4, the graph catalogs the non-synonymous mutations found in the exonic regions of the genes of interest, which were not previously listed as variants of interest by the user.

The modularity of the NeST framework and standardization of results generated from the framework allow users not only to identify high-quality variants but also allow for the benchmarking of commonly used algorithms for variant calling. In the next section, the accuracy of standard variant calling algorithms using *in-silico* datasets from *Plasmodium* genes is evaluated.

4.4 *In-silico* evaluation of variant calling accuracy from NGS datasets

4.4.1 Materials and methods:

To evaluate the accuracy of each methodology implemented in NeST and added value of consensus variant calling, 108 *in-silico* MiSeq dataset with varying coverage and sequence error rates were compared. Complete gene sequences for *PfCRT, PfMDR1, PfK13, PfDHPS, PfDHFR,* and the complete mitochondrial genome from *Plasmodium falciparum 3D7* genome were used as the reference. The mutation rate for the *in-silico* samples was set to 0.001%, with 10% of the in-silico mutations generated as InDels using DWGSIM[90] as described in Table 4.1. The specificity and sensitivity of each variant caller was assessed using the variant call files generated by DWGSIM as a truth set.

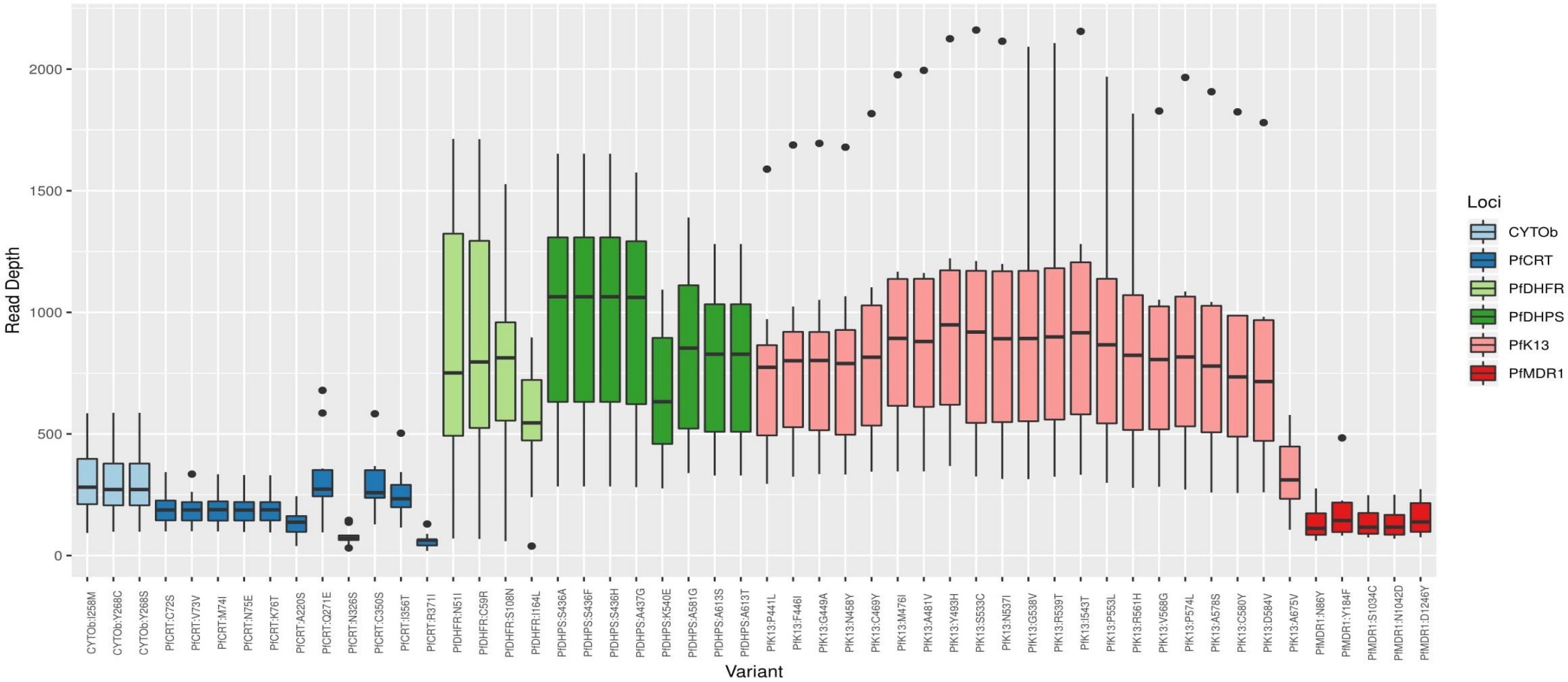


Figure 4.3: Depth of sequencing coverage across regions of interest. The y-axis indicates the sequencing read depth; the x-axis lists the variants of interest as specified by the user. Here, the variants of interest are mutations conferring antimalarial drug resistance**.**

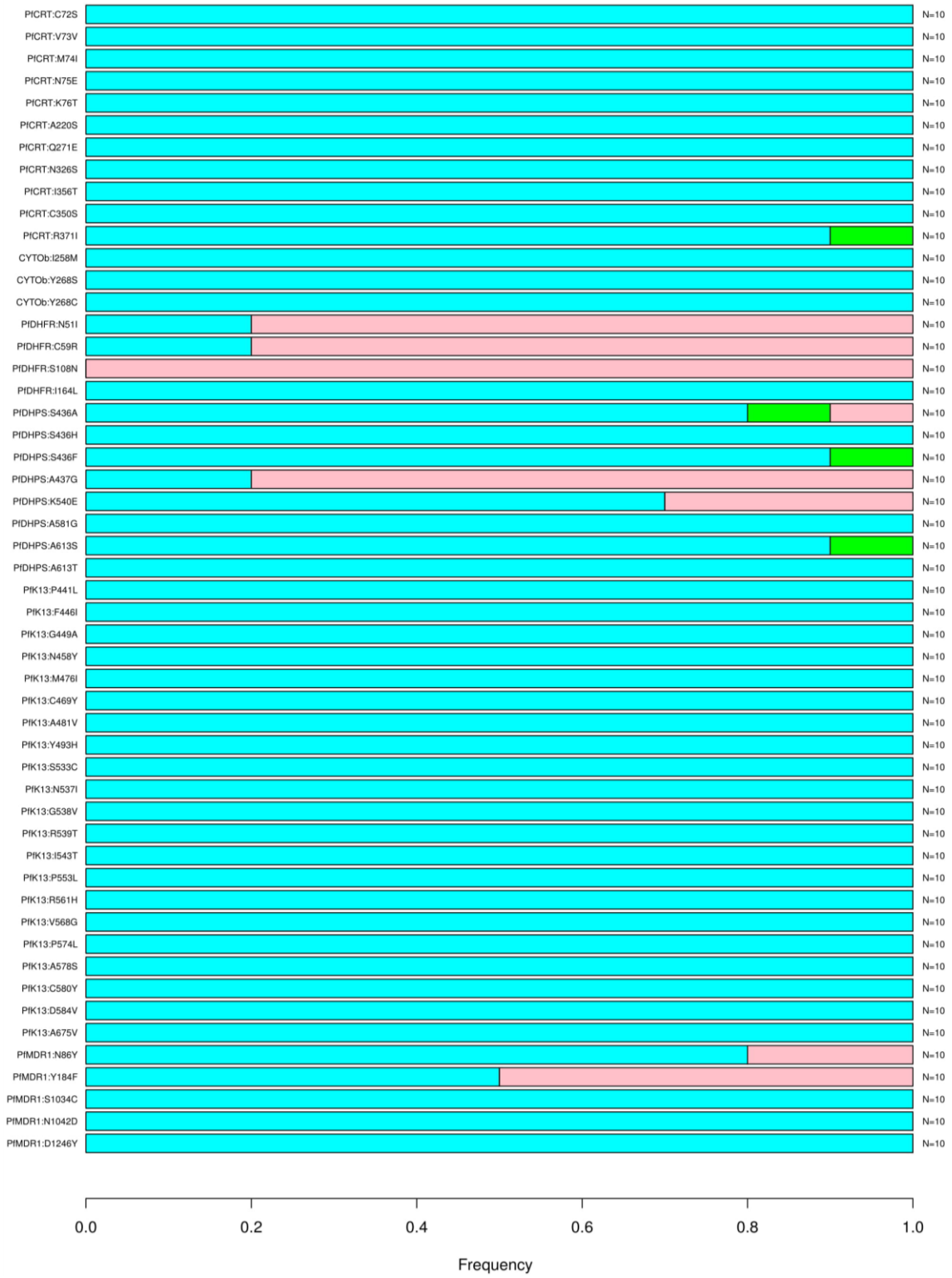


Figure 4.4: Allele frequency distribution across variant of interest. The y-axis lists variants of interest as specified by the user. The x-axis indicates the allele frequency in the sample set. The color of the indicate the allele balance for the variant. Blue indicates wild type, green indicates variant in the minor allele, and red indicates presence of major allele variant.

Table 4.4: Characteristics of *in-silico* amplicon data set generated.

Set#	Coverage	Read count	Error rate	MAF
1	100	4010	0.001	0.5
2	500	20051	0.001	0.5
3	1000	40102	0.001	0.5
4	100	4010	0.001	0.25
5	500	20051	0.001	0.25
6	1000	40102	0.001	0.25
7	100	4010	0.001	0.1
8	500	20051	0.001	0.1
9	1000	40102	0.001	0.1
10	100	4010	0.001	0.05
11	500	20051	0.001	0.05
12	1000	40102	0.001	0.05
13	100	4010	0.005	0.5
14	500	20051	0.005	0.5
15	1000	40102	0.005	0.5
16	100	4010	0.005	0.25
17	500	20051	0.005	0.25
18	1000	40102	0.005	0.25
19	100	4010	0.005	0.1
20	500	20051	0.005	0.1
21	1000	40102	0.005	0.1
22	100	4010	0.005	0.05
23	500	20051	0.005	0.05
24	1000	40102	0.005	0.05
25	100	4010	0.01	0.5
26	500	20051	0.01	0.5
27	1000	40102	0.01	0.5
28	100	4010	0.01	0.25
29	500	20051	0.01	0.25
30	1000	40102	0.01	0.25
31	100	4010	0.01	0.1
32	500	20051	0.01	0.1
33	1000	40102	0.01	0.1
34	100	4010	0.01	0.05
35	500	20051	0.01	0.05
36	1000	40102	0.01	0.05

4.4.2 Results:

The accuracy of the different variant calling algorithms was evaluated using, 108 paired-ended *in-silico* datasets were following the coverage, error rate, and minor allele frequency combination listed in Table 4.4. The maximum error rate for samples generated was limited to 1% since any error rate higher than 1% resulted in all reads being discarded during the QC step, thus not resembling any real-world situation which would be encountered. For each of the samples the mutation rate was set to 0.001% and rate of InDels was set to 10% of all mutations simulated.

Table 4.5: Precision and recall values for SNPs and InDel calls made by standard variant callers against *in-silico* datasets from *Plasmodium falciparum* genes.

Variant caller	All variant calls (2241)			SNPs (2043)			InDels (198)		
	Precision	Recall	Support	Precision	Recall	Support	Precision	Recall	Support
NeST	61.88	87.50	3169	63.36	94.66	3052	23.07	13.63	117
NeST (Conf=2)	99.88	77.10	1730	99.94	84.53	1728	50	0.50	2
HaplotypeCaller	98.04	78.40	1792	99.94	84.67	1731	44.26	13.63	61
Samtools	97.83	70.72	1620	99.81	77.53	1587	3.03	0.55	33
Freebayes	62.72	85.58	3058	63.23	93.88	3033	0	0	25

The variant calls made with NeST, Freebayes, Samtools, GATK HaplotypeCaller and NeST calls which were made by at least two of three variant callers present in the framework were compared with the mutation list generated for each sample by DWGSIM[90]. Precision and Recall was calculated by measuring the number of instances where the calls made through the different protocols exact matched in-silico mutations generated.

Table 4.5 shows the precision and recall of the different protocols broken down by the type of variant call. From the table, we can see that relying on the simple comparison

of InDel calls to arrive at a consensus variant call is not a viable option. Primarily due to the differences in the representation of InDels by different variant calling algorithms. Precision values for InDel calling are greatly improved if InDels are unfurled by base location and then compared, as suggested by *Krusche et al.*[103]. The current version of NeST is primarily aimed at identifying SNPs conferring antimalarial drug resistance. InDels are just reported as is, giving the user the option to consider InDels for their downstream analysis.

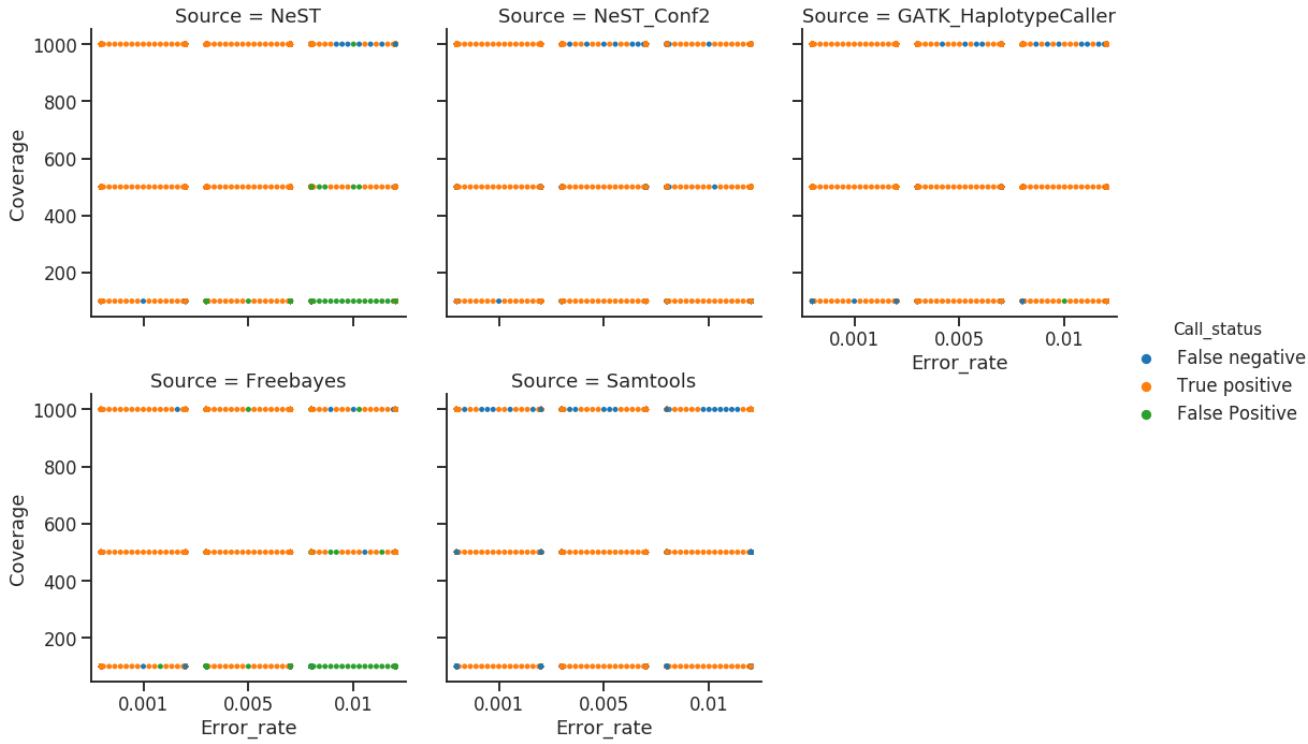


Figure 4.5: Distribution of false positive and false negative calls made by the different variant calling methods as a function of sequencing depth and error rates. The swarm plot shows the density of false positive and false negative calls for each condition.

With SNP calling, however, HaplotypeCaller and Samtools show high precision and recall. Freebayes, on the other hand, shows significantly lower precision but high recall when compared to the other two callers, as well as NeST (Confidence =2). To investigate

this further, the variance in error in the variant calls w.r.t coverage, sequence error rate, and MAF were assessed. From Figure 4.5, the precision of Freebayes is affected to a more significant extent by sequence error rate than the other methods implemented. It is interesting to see that with an increase in coverage, the number of false-positive calls made by Freebayes significantly decreases. Samtools and GATK seem to take a more conservative approach while calling variants, thus showing greater precision, and consistent recall across the different scenarios described here.

From the data it's clear that NeST results filtered on at least two variant callers detecting the variant, shows higher precision than using any of the methods on their own. When it comes to consensus InDel calling, the biggest bottleneck is the differing representations of InDels by the different methods implemented. Though decomposing the InDel into its constituent bases can enable consensus InDel calling, the current implementation of NeST reports InDels in the native forms as called by the different variant caller.

4.5 Identifying variants conferring antimalarial drug resistance in *Plasmodium falciparum* from Targeted Amplicon Deep Sequencing datasets

4.5.1 Materials and methods

Amplicon sequencing is a targeted sequencing approach where PCR products or amplicons from a specific region in the genome are sequenced to a very high depth of coverage. This enables the accurate characterization of the genomic variants. Multiplexing of these amplicons allows for a large number of samples to be sequenced in a single sequencing run. Using the targeted amplicon deep sequencing protocol developed for eight genetic

markers for drug resistance in *Plasmodium falciparum*[48], NeST was used to identify variants associated with drug resistance from these samples.

A total of 1,081 *Plasmodium falciparum* samples from more than 28 different regions (NCBI BioProject PRJNA428490) have been sequenced till date. From this larger dataset, 243 isolates were sequenced using Sanger sequencing (Figure 4.1). Using these 243 samples the accuracy of the variant calls from NeST were evaluated by comparing the Sanger and NGS variant calls.

The accuracy of NeST was determined by its ability to identify 29 SNPs associated with drug resistance that were amplified in the Sanger sequencing runs. The results from NeST were compared with variant calls from Geneious[98], a commercial off the shelf (COTS) bioinformatics toolkit for Sanger and NGS data analysis.

Leveraging the HPC compatibility of NeST, the framework was made available to all the groups at the CDC through their internal HPC framework. In collaboration with the Office of Advanced Molecular Detection (OAMD) at the CDC, a cloud-based web version of NeST was deployed on the CDC OAMD portal, accessible by all collaborating public health laboratories across the world (Figure 4.10).

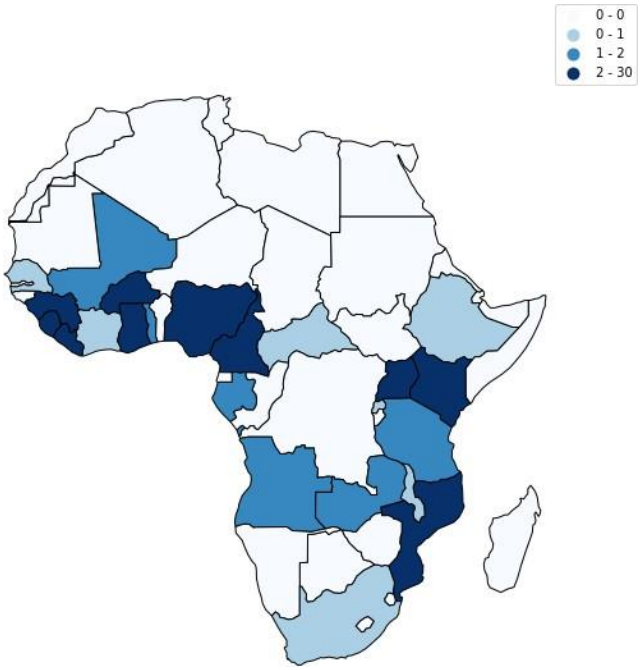


Figure 4.6: Summary of the geographical location of all 243 *P. falciparum* samples. Legend indicates the number of samples from each region.

4.5.2 Results

The Malaria Resistance Surveillance (MaRS) amplicon sequencing protocol developed by the Malaria Branch at the CDC currently uses PCR primers to amplify out whole gene sequences for these eight markers (*PfCRT, PfMDR1, PfDHPS, PfDHFR, PfK13, PfCOXIII, PfCOL,* and *PfCYTb*). These eight markers were amplified in the 243 samples of imported malaria cases into the United States, using the MaRS protocol. NeST was used to identify mutations conferring drug resistance. The results were validated using variant calls from Sanger sequencing data available for all 243 samples.

As indicated in Figure 4.7, NeST was able to detect 444 variants in the 243 samples, that were previously missed by Sanger sequencing and Geneious. Six hundred three

variants were only detected through NGS data. When compared to Geneious, NeST and Sanger were able to identify 129 SNPs that were missed by Geneious. Only 21 variants from all 243 samples were missed by NeST but detected either with Sanger (20 variants) or Geneious.

Figure 4.8 shows the overlap of variant calls made by the different tools used within NeST. From the figure we see that 1566 variant calls detected from the Sanger, 1543 were found by at least two variant callers implemented in NeST. Of 603 variants only identified from NGS data, 577 were identified by at least two variant callers in NeST. Freebayes detected the largest number of variant calls that could not be corroborated with the other methods.

Samtools detected the greatest number of variant calls, which could only be corroborated with sanger data. All variant calls made by GATK, on the other hand, could be verified by at least one of the variant callers used in the comparison or sanger sequencing data. Thus, GATK again shows that while being conservative in the number of calls made, the accuracy of the calls made by the tools is the highest amongst the tools compared here.

Apart from the variants of interest, any other variant called by NeST will be stored in separate outputs as novel intronic or novel exonic variants while novel exonic variants are only reported if at least two out of the three callers found the variant. Novel intronic calls are reported even when just one of the callers found the variant. This separation and subsequent reporting of novel variants can help with the discovery of potential markers for drug resistance.

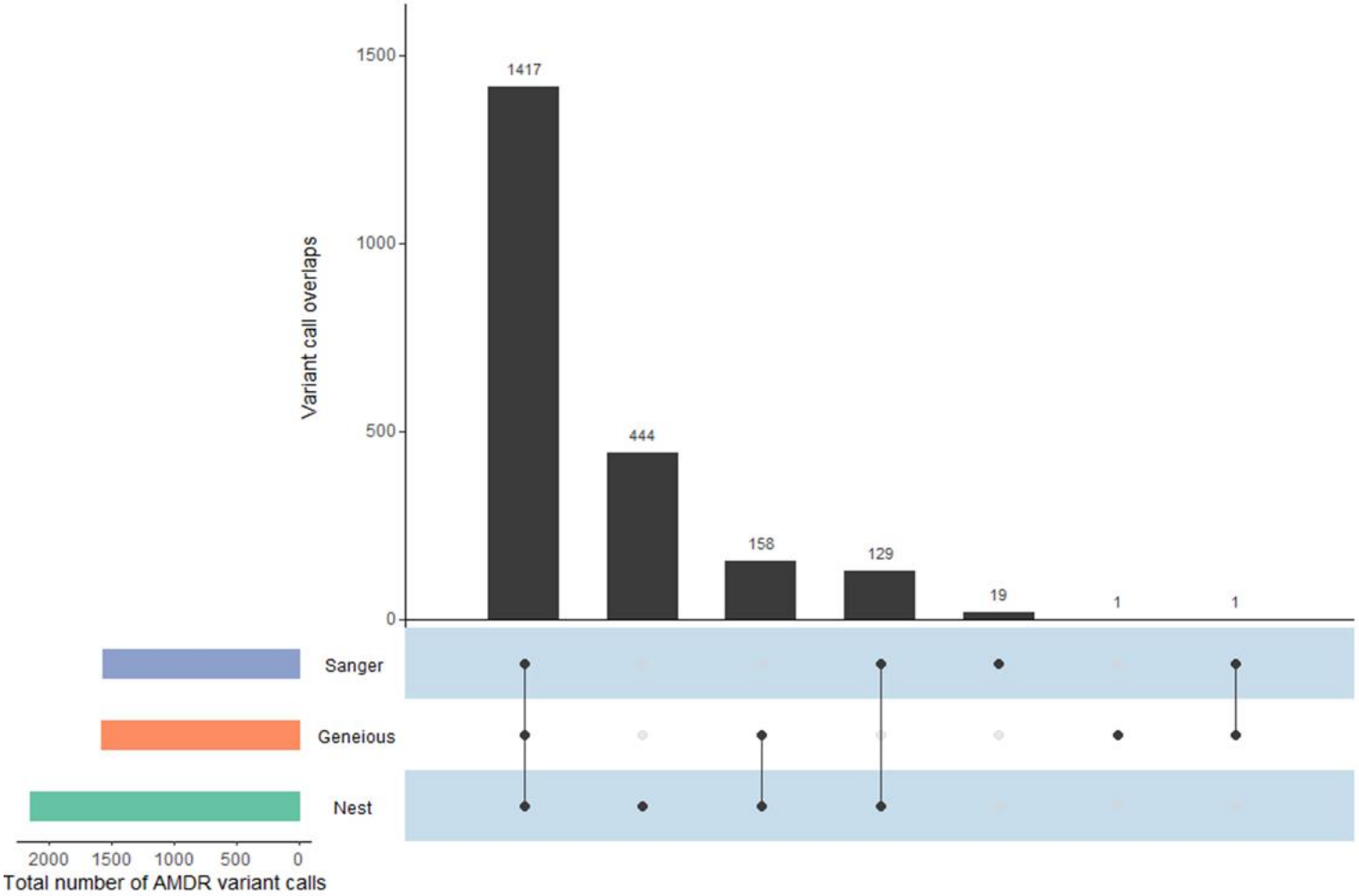


Figure 4.7: The overlap of variant calls made by NeST, Geneious, and Sanger sequencing calls. The bar graph on the left shows the total number of variant calls made by the different methods. The graph on the right shows the extent of overlap of variant calls made by NeST, Geneious, and Sanger calls.

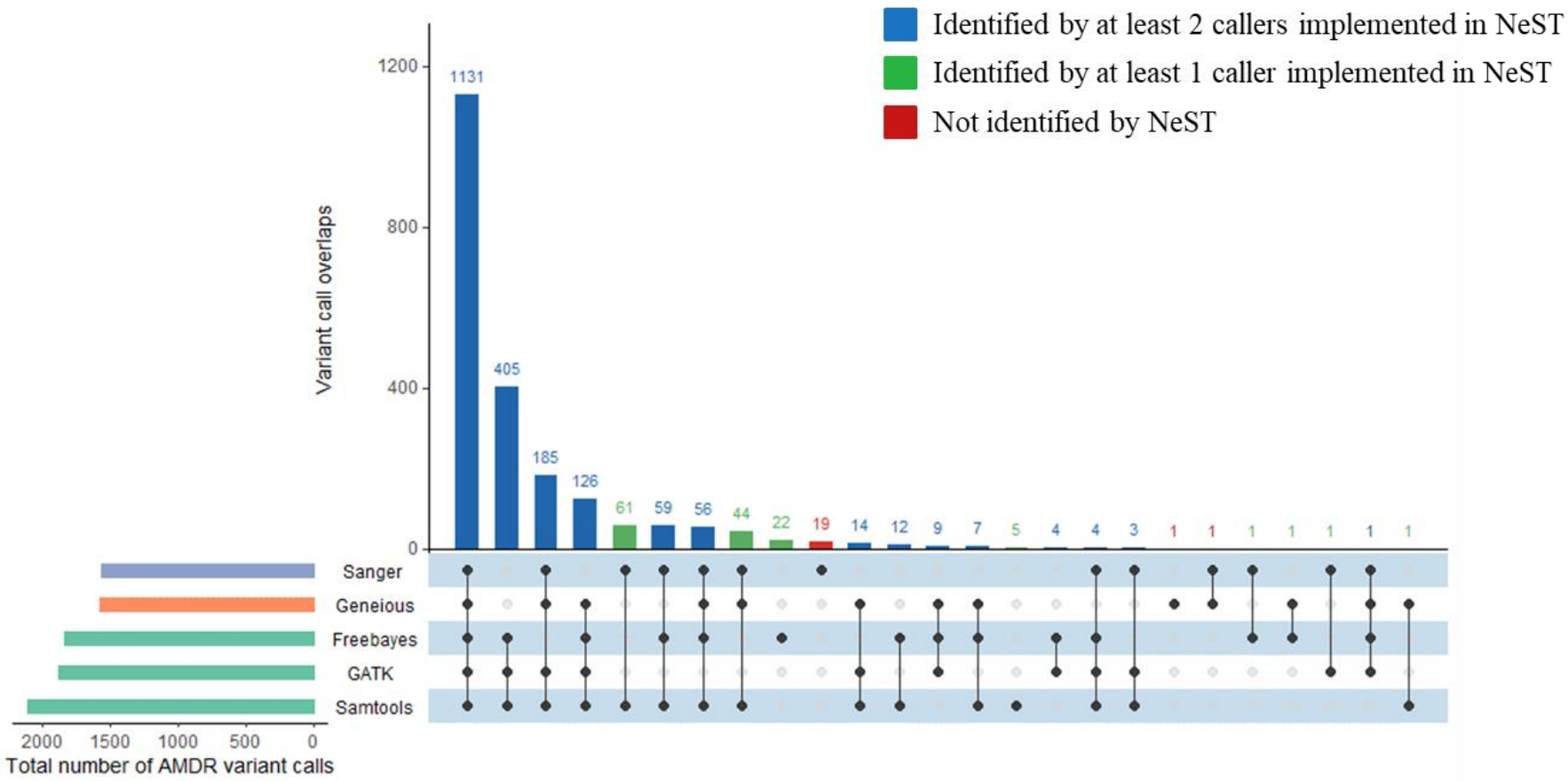


Figure 4.8: The overlap of variant calls made by NeST, Geneious, and Sanger sequencing calls. The bar graph on the left shows the total number of variant calls made by the different methods. The graph on the right shows the extent of overlap of variant calls made by Samtools, GATK HaplotypeCaller, Freebayes, Geneious, and Sanger calls. Vertical bars are colored by the methods that identified the variant.

.

Figure 4.9 shows novel non-synonymous exonic mutations in the *PfK13* gene. From the figure, we can see that 115 of the 243 samples analyzed had the *PfK13*: K189T mutation. While it is difficult to arrive at any conclusion about the significance of this mutation toward drug resistance with the current sample set. Capturing this information can help with future inquiries.

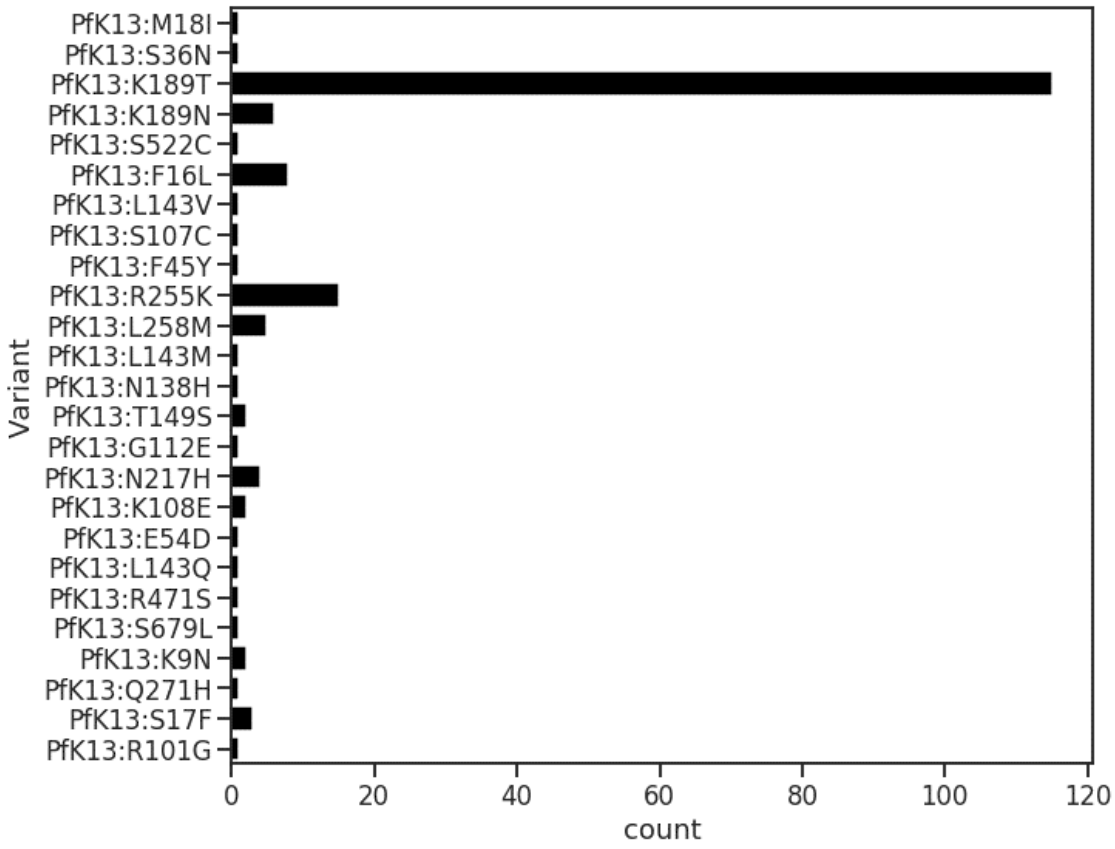


Figure 4.9: Frequency of novel non-synonymous mutations in the exonic region *PfK13* that were found by at least two variant callers.

Our analysis of the 243 *P. falciparum* isolates shows the advantage of using a consensus-based variant calling system like NeST for tracking resistance in malaria. To ensure that NeST can be deployed quickly across different platforms with varying resources available, NeST using the BioConda framework to setup necessary dependencies. It is also structured

such that the user requires a minimum of 4GB of RAM for analysis, which is commonplace in most modern-day computers. For situations where adequate resources to run NeST are not available, a web-based version of the framework was also made available to all collaborating public health labs at CDC, through the Office of Advanced Molecular Detection (OAMD) web-portal (Figure 4.10) as the MaRS pipeline.

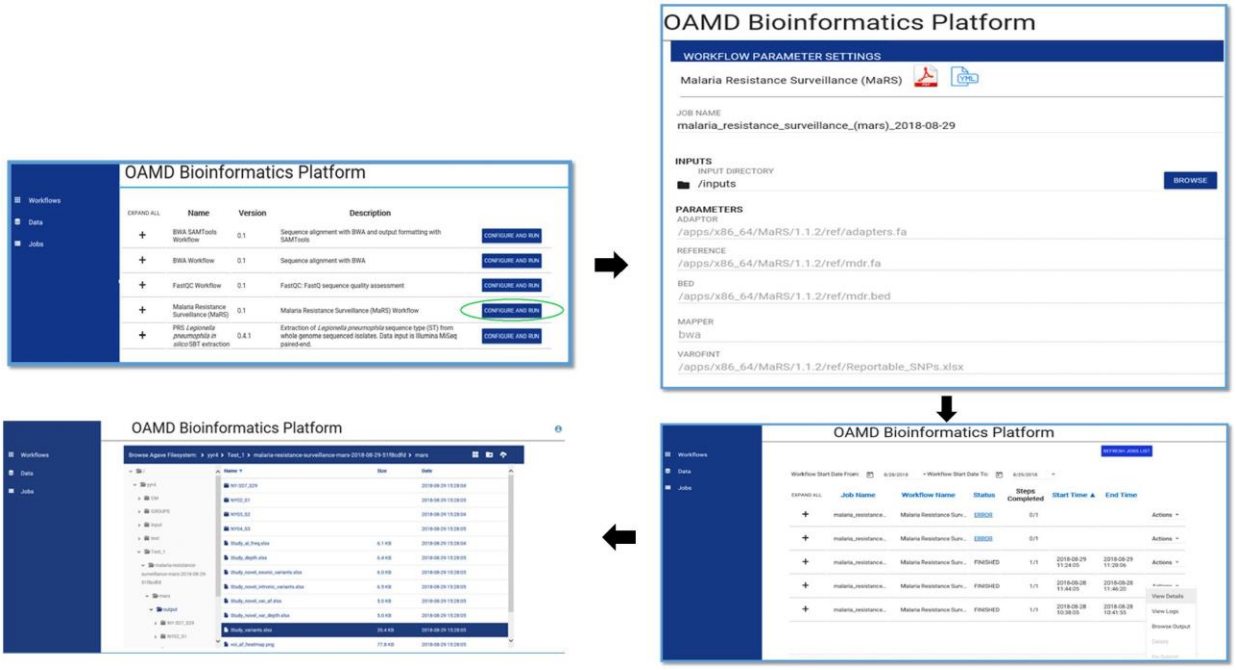


Figure 4.10: NeST/ MaRS app available on the OAMD Bioinformatics platform.

4.6 Evaluating accuracy of variant calling derived genotypes at predicting phenotypic drug resistance in *Mycobacterium tuberculosis*

4.6.1 Material and methods

To test the utility of NeST for a bacterial pathogen, the performance of NeST at predicting drug resistance in *Mycobacterium tuberculosis* isolates from the CRyPTIC consortium

study[83] was evaluated. From the CRyPTIC consortium study, 8351 samples along with their corresponding SRA accession numbers were collected. NeST was used to identify mutations conferring drug resistance to four drugs Rifampicin, Isoniazid, Ethambutol, and Pyrazinamide. We looked at mutations in 9 genes *katG, embA, embB, embC, ahpC, pncA, inhA, fabG1* and *rpoB*.

The study presented 10210 *Mycobacterium tuberculosis* isolates with associated phenotypic resistance information for five anti-biotics lists of anti-biotics. They tested the accuracy of genotypic variant calling markers from NGS datasets at predicting phenotypic resistance by implementing a Standard variant calling a pipeline. However, of the 10210, only 8351 samples available from NCBI's Sequence Read Archive (SRA) could be associated with the phenotypic associations. Only these 8351 samples were considered for further analysis.

Here we evaluate the accuracy of NeST at predicting phenotypic resistance to these drugs using resistance variant markers for these five drugs. The sensitivity and specificity of predicting phenotypic resistance using a single variant caller as performed in the CRyPTIC consortium study and a consensus variant calling approach implemented in NeST, was tested. Samples were analyzed using the NeST variant calling framework and classified as resistant or susceptible based on presence or absence of variants known to confer drug resistance to four first line treatments against *M. tuberculosis* infections.

4.6.2 Results:

To test the utility of NeST in different public health settings, we evaluated the performance of NeST at predicting drug resistance in *Mycobacterium tuberculosis* isolates from the

CRyPTIC consortium study[83]. From the CRyPTIC consortium study, we were able to gather samples and their corresponding SRA accession numbers for 8351 samples. We used to NeST to identify mutations conferring drug resistance to four drugs Rifampicin, Isoniazid, Ethambutol, and Pyrazinamide. In particular, we looked at mutations in 9 genes *katG, embA, embB, embC, ahpC, pncA, inhA, fabG1* and *rpoB*.

The current implementation of NeST leverages the Python multiprocessing framework. This allows the distribution of the analysis across multiple threads to achieve the parallelization of the process. While multiprocessing is suitable for small studies, as we saw with the *Plasmodium* dataset, it is not efficient when working with a large number of samples. Leveraging the multi-node structure of an HPC environment can drastically speed up the analysis. To this end, we developed a recipe for NeST to utilize the multi-node architecture of the HPC system, PACE[92], offered at the Georgia Institute of Technology. Since the MOAB scheduler used by PACE, is a commonly used scheduler for HPC clusters, the same recipe can be used across any cluster implementing a MOAB scheduler. Leveraging this architecture, we were able to analyze the 8351 WGS samples in 40 hours. The samples were classified as resistant to the drugs if an SNP known to confer drug resistance was found in the sample.

For comparison, the convention set by the published study of classifying a sample to be susceptible to drugs was followed; if it lacked any of the SNPs known to confer drug resistance. Precision and recall values were calculated for each of the labels (Resistant, Susceptible). Table 4.6 shows the conditions under which a true positive, false positive and false negative call would be made for both the resistant and susceptible label.

Table 4.6: Truth table for genotypic and phenotypic resistance or susceptibility.

	Genotypic resistance	Genotypic susceptibility
Phenotypic resistance	True positive resistance	False negative resistance False positive susceptibility
Phenotypic susceptibility	False negative resistance False positive susceptibility	True positive susceptibility

$$Precision = \frac{True\ Positive}{True\ Positive \cup False\ Positive} \quad (4.2)$$

$$Recall = \frac{True\ Positive}{True\ Positive \cup False\ Negative} \quad (4.2)$$

As per the definition of precision and recall, Eq (4.1) (4.2), low precision for resistance would imply that more samples were predicted to be resistant from the NGS data, than what was observed from the clinical phenotype. A low value for recall would imply fewer samples were predicted to be resistant from the NGS data, than what was observed from clinical phenotype. By the same logic, a low precision value corresponding to susceptibility to a drug would imply that there were a greater number of samples that were labeled susceptible than observed through phenotypic data. Moreover, low values of recall would indicate that a greater number of samples were labeled resistant than expected as per the phenotypic information.

Table 4.7: Precision and recall values for the genotypic prediction of phenotypic resistance or susceptibility against anti-TB drugs.

Drug	Label	NeST		CRyPTIC consortium		Support
		Precision	Recall	Precision	Recall	
Rifampicin	Resistant	0.94	0.96	0.96	0.95	2659
	Susceptible	0.98	0.97	0.98	0.98	5692
Pyrazinamide	Resistant	0.46	0.43	0.59	0.75	1016
	Susceptible	0.92	0.93	0.96	0.93	7335
Ethambutol	Resistant	0.66	0.83	0.66	0.86	1488
	Susceptible	0.96	0.91	0.97	0.91	6863
Isoniazid	Resistant	0.93	0.93	0.94	0.93	2884
	Susceptible	0.97	0.96	0.96	0.97	5467

Table 4.7 shows the precision and recall corresponding to resistant and susceptible calls from NEST and the CRyPTIC consortium pipeline. Based on the interpretations of precision and recall for the different scenarios, we can say that NeST classifies a lot more samples as resistant to the four drugs, than the Samtools pipeline used by the CRyPTIC consortium. However, it would be unwise to conclude that NeST is better at detecting phenotypic resistance than a single pipeline method or vice versa. While from the previous sections we see that the consensus framework can achieve higher accuracy of variant calls, there are many factors involved with phenotypic resistance than cannot be accounted for by a naive classification of resistance based on presence or absence of variants in the samples.

From the data provided in Table 4.7, however, the precision and recall associated with the phenotypic prediction of susceptibility to a drug was above 90% for both CRyPTIC consortium pipeline and NeST. It might be tempting to conclude here that,

absence of the mutation associated with resistance can imply phenotypic susceptibility. The absence of a variant call from NGS data is dependent on many factors. Such as the purity of DNA extracted, the precision of the library prep, the error rates associated with the sequencing technology used, and finally, the sensitivity of the NGS analysis framework used. More evidence is needed before any conclusion can be made on the accuracy of genotypic predictions of phenotypic resistant. Based on the analysis presented herein, NeST provides a scalable framework to analyze NGS data from large-scale studies. The consensus framework offered provides a metric to assess high confidence variant calls. These two features can significantly improve the utilization of NGS data towards molecular surveillance of drug resistance.

4.7 Discussion

Though variant calling from NGS data has been an established technique for some time now, there is ample evidence to show that variant calls made by different tools are not always consistent with each other[63,65]. From the analysis using *in-silico* datasets generated from genes associated with drug resistance in *Plasmodium falciparum*, its clear that variance in results will occur with different variant calling algorithms and that some of this variance can be overcome by using a consensus based variant calling method.

NeST is already being used for routine molecular surveillance of antimalarial drug resistance at the CDC. Based on the analysis presented earlier, NeST also performs better than commercial toolkits such as Geneious at identifying genetic markers for antimalarial drug resistance. The scalability of NeST is demonstrated by its availability through the

CDC HPC web portal. Furthermore, NeST is easily scalable for large datasets, as demonstrated by the analysis of the 8,351 NGS samples from the CRyPTIC *Mycobacterium tuberculosis* consortium study. While most of the variant calls associated with drug resistance were able to predict phenotypic resistance to a drug, some drugs, such as Pyrazinamide and Ethambutol, the SNP calls were not good predictor of phenotypic resistance. The SNP calls predicted a lot more samples to be resistant to a drug than phenotypically observed. This may be due to the various factors that can affect variant calling from NGS data, such as coverage, error rate, and genomic complexity. Additionally, the mechanisms that govern resistance to a drug are complex and thus are not easily predicted by naive presence or absence test of variants associated with drug resistance – genotypic resistance is not phenotypic resistance and vice versa.

In summary, the consensus-based variant calling framework can overcome the biases of any individual bioinformatic algorithm and provides a metric to assess the quality of mutations detected from NGS datasets. Though by its nature, a consensus-based system will be more conservative in calling variations from the sample data, consensus-based methods can reduce the likelihood of erroneous calls, which is crucial for molecular surveillance in public health settings. NeST was able to identify more variants associated with drug resistance, than other commercial NGS analysis platforms. The scalability and modular nature of NeST make it an ideal tool for the large-scale molecular surveillance of drug resistance.

CHAPTER 5. CONCLUDING REMARKS

The work described in this thesis highlights the challenges faced in genetic epidemiology of malaria and proposes methods to address these challenges given the current technology.

In Chapter 2, we describe the existing methodologies for variant calling and genomic clustering from NGS data. The advantages and disadvantages of state-of-the-art genomic clustering and variant calling algorithms were discussed. The underlying principle for two new methodologies, NeST and Gentoo, introduced through this work to overcome the pitfalls of existing solutions in variant calling and genomic clustering when applied to NGS data in a public health setting.

In Chapter 3, I further expanded on the k-mer based alignment-free clustering algorithm, Gentoo. The improved accuracy of Gentoo over existing methods for genomic clustering of NGS data from *Plasmodium* spp., was demonstrated. I further showed that Gentoo was able to cluster better genomes from *Candida auris* outbreaks in Colombia, based on local similarities between isolates from a given outbreak, when compared to SNP based phylogeny techniques.

Memory profiling of Gentoo shows a significant reduction in the time of execution and resources required for pairwise distance estimation that can be achieved with Gentoo. Gentoo required 121 MB of RAM and took 10 minutes to cluster all 20 genomes from the *Plasmodium* study. While our previously developed method, Finch, required 10 GB of RAM and took 30 minutes to perform all 210 pairwise comparisons. The time of analysis went up to 60 minutes when Gentoo used FASTQ files directly, but the memory utilization

remained the same. While Mash was the most efficient algorithm, in terms of runtime and memory utilization, Gentoo was more accurate at reproducing the *Plasmodium* evolutionary tree.

Chapter 4 describes a modular, scalable, consensus-based framework, NeST, for the molecular surveillance of drug resistance using variant calling from NGS datasets. There has been enough evidence to show that variant calling methodologies are not consistent with each other, though methods do exist to filter low-quality variants. These variant filtration techniques are not applicable for smaller organisms such as *Plasmodium falciparum* or *Mycobacterium. tuberculosis.*

NeST generates a consensus call using multiple variant calling algorithms to overcome the potential bias associated with each method on its own. We evaluated the benefits of a consensus caller, using *in-silico* datasets generated from molecular markers associated with drug resistance in *P. falciparum* as well as the amplicon sequencing data from blood spots of samples infected with *P. falciparum*. We demonstrated the scalability of the framework by analyzing 8351 *M. tuberculosis* on an HPC cluster, with an average time of analysis of 4 hours for 100 WGS samples of *M. tuberculosis*.

We show that there is a consistency in the precision when calling variants using a consensus-based approach as compared to each method separately. However, there is drop in the recall, up to 8% drop when compared to the variant caller with the highest recall. The trade-off between precision and recall, however, needs to be evaluated on a case-by-case basis. We offer a confidence metric for each variant call that assists in decision making downstream.

Future versions of NeST, however, need to address the issue of consensus InDel calling. The main bottleneck with regard is the difference in the representation of InDels by the various methods. One possible solution discussed in the previous section was unfurling InDel calls and locally assembling the consensus calls, the computational burden of this proposed solution needs to be evaluated further.

Through my Ph.D. thesis work, I aimed to advance the utilization of NGS technology in the public health sector. The solutions offered in this thesis provide an efficient and scalable framework for genomic clustering and molecular surveillance of drug resistance in malaria. The utility of NGS analysis methods that I developed is demonstrated by their adaptation as standard protocols by the Malaria Branch at the Centers for Disease Control and Prevention. Considering the standardization and improved accuracy, I hope that public health labs across the world adopt the solutions described in the present work for the surveillance of other infectious diseases.

REFERENCES

1. Health Organization, W. *WORLD MALARIA REPORT 2018 ISBN 978 92 4 156565 3*.

2. Cowman, A. F., Healer, J., Marapana, D. & Marsh, K. Malaria: Biology and Disease. *Cell* **167**, 610–624 (2016).

3. Meibalan, E. & Marti, M. Biology of Malaria Transmission. *Cold Spring Harb. Perspect. Med.* **7**, a025452 (2017).

4. Blasco, B., Leroy, D. & Fidock, D. A. Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. *Nat. Med.* **23**, 917–928 (2017).

5. Menard, D. & Dondorp, A. Antimalarial Drug Resistance: A Threat to Malaria Elimination. *Cold Spring Harb. Perspect. Med.* **7**, a025619 (2017).

6. Protopopoff, N. *et al.* Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two fact. *Lancet (London, England)* **391**, 1577–1588 (2018).

7. Masaninga, F. *et al.* Insecticide-treated nets mass distribution campaign: benefits and lessons in Zambia. *Malar. J.* **17**, 173 (2018).

8. Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. *Nat. Biotechnol.* **34**, 78–83 (2016).

9. Kyrou, K. *et al.* A CRISPR–Cas9 gene drive targeting doublesex causes complete population suppression in caged Anopheles gambiae mosquitoes. *Nat. Biotechnol.* **36**, 1062–1066 (2018).

10. Gantz, V. M. *et al.* Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. *Proc. Natl. Acad. Sci.* **112**, E6736--E6743 (2015).

11. Conklin, B. R. On the road to a gene drive in mammals. *Nature* **566**, 43–45 (2019).

12. Draper, S. J. *et al.* Malaria Vaccines: Recent Advances and New Horizons. *Cell Host Microbe* **24**, 43–56 (2018).

13. Warhurst, D. C. & Williams, J. E. ACP Broadsheet no 148. July 1996. Laboratory diagnosis of malaria. *J. Clin. Pathol.* **49**, 533–8 (1996).

14. Moody, A. Rapid diagnostic tests for malaria parasites. *Clin. Microbiol. Rev.* **15**, 66–78 (2002).

15. Haldar, K., Bhattacharjee, S. & Safeukui, I. Drug resistance in Plasmodium. *Nat. Rev. Microbiol.* **16**, 156–170 (2018).

16. Noedl, H. *et al.* Evidence of Artemisinin-Resistant Malaria in Western Cambodia. *N. Engl. J. Med.* **359**, 2619–2620 (2008).

17. Suresh, N. & Haldar, K. Mechanisms of artemisinin resistance in Plasmodium falciparum malaria. *Curr. Opin. Pharmacol.* **42**, 46–54 (2018).

18. *METHODS FOR SURVEILLANCE OF ANTIMALARIAL DRUG EFFICACY WHO Library Cataloguing-in-Publication Data : Methods for surveillance of antimalarial drug efficacy*. (2009).

19. Djimdé, A. *et al.* A Molecular Marker for Chloroquine-Resistant Falciparum Malaria. *N. Engl. J. Med.* **344**, 257–263 (2001).

20. Sidhu, A. B. S., Verdier-Pinard, D. & Fidock, D. A. Chloroquine resistance in Plasmodium falciparum malaria parasites conferred by pfcrt mutations. *Science* **298**, 210–213 (2002).

21. Nkrumah, L. J. *et al.* Probing the multifactorial basis of Plasmodium falciparum quinine resistance: Evidence for a strain-specific contribution of the sodium-proton exchanger PfNHE. *Mol. Biochem. Parasitol.* **165**, 122–131 (2009).

22. Cooper, R. A. *et al.* Mutations in transmembrane domains 1, 4 and 9 of the Plasmodium falciparum chloroquine resistance transporter alter susceptibility to chloroquine, quinine and quinidine. *Mol. Microbiol.* **63**, 270–282 (2007).

23. Petersen, I. *et al.* Balancing drug resistance and growth rates via compensatory mutations in the *P lasmodium falciparum* chloroquine resistance transporter. *Mol. Microbiol.* **97**, 381–395 (2015).

24. Pascual, A. *et al.* In vitro piperaquine susceptibility is not associated with the

Plasmodium falciparum chloroquine resistance transporter gene. *Malar. J.* **12**, 431 (2013).

25. Dhingra, S. K. *et al.* A Variant PfCRT Isoform Can Contribute to Plasmodium falciparum Resistance to the First-Line Partner Drug Piperaquine. *MBio* **8**, e00303--17 (2017).

26. Shah, N. K. *et al.* Antimalarial drug resistance of Plasmodium falciparum in India: changes over time and space. *Lancet Infect. Dis.* **11**, 57–64 (2011).

27. Tumwebaze, P. *et al.* Changing antimalarial drug resistance patterns identified by surveillance at three sites in Uganda. *J. Infect. Dis.* **215**, jiw614 (2016).

28. Costa, G. L. *et al.* Assessment of copy number variation in genes related to drug resistance in Plasmodium vivax and Plasmodium falciparum isolates from the Brazilian Amazon and a systematic review of the literature. *Malar. J.* **16**, 152 (2017).

29. Peterson, D. S., Walliker, D. & Wellems, T. E. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 9114–9118 (1988).

30. Gregson, A. & Plowe, C. V. Mechanisms of resistance of malaria parasites to antifolates. *Pharmacol. Rev.* **57**, 117–45 (2005).

31. Korsinczky, M. *et al.* Mutations in Plasmodium falciparum cytochrome b that are associated with atovaquone resistance are located at a putative drug-binding site.

Antimicrob. Agents Chemother. **44**, 2100–2108 (2000).

32. Srivastava, I. K. & Vaidya, A. B. A mechanism for the synergistic antimalarial action of atovaquone and proguanil. *Antimicrob. Agents Chemother.* **43**, 1334–1339 (1999).

33. Goodman, C. D., Buchanan, H. D. & McFadden, G. I. Is the Mitochondrion a Good Malaria Drug Target? *Trends Parasitol.* **33**, 185–193 (2017).

34. Cowman, A. F., Galatis, D. & Thompson, J. K. Selection for mefloquine resistance in Plasmodium falciparum is linked to amplification of the pfmdr1 gene and cross-resistance to halofantrine and quinine. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1143–1147 (1994).

35. Pickard, A. L. *et al.* Resistance to antimalarials in Southeast Asia and genetic polymorphisms in pfmdr1. *Antimicrob. Agents Chemother.* **47**, 2418–2423 (2003).

36. Stepniewska, K. *et al.* In vivo assessment of drug efficacy against Plasmodium falciparum malaria: duration of follow-up. *Antimicrob. Agents Chemother.* **48**, 4271–4280 (2004).

37. Sidhu, A. B. S., Valderramos, S. G. & Fidock, D. A. pfmdr1 mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in Plasmodium falciparum. *Mol. Microbiol.* **57**, 913–926 (2005).

38. Reed, M. B., Saliba, K. J., Caruana, S. R., Kirk, K. & Cowman, A. F. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium

falciparum. *Nature* **403**, 906–909 (2000).

39. Duraisingh, M. T., Roper, C., Walliker, D. & Warhurst, D. C. Increased sensitivity to the antimalarials mefloquine and artemisinin is conferred by mutations in the pfmdr1 gene of Plasmodium falciparum. *Mol. Microbiol.* **36**, 955–961 (2000).

40. Mwai, L. *et al.* In vitro activities of piperaquine, lumefantrine, and dihydroartemisinin in Kenyan Plasmodium falciparum isolates and polymorphisms in pfcrt and pfmdr1. *Antimicrob. Agents Chemother.* **53**, 5069–5073 (2009).

41. Health Organization, W. *Artemisinin resistance and artemisinin-based combination therapy efficacy (Status report -- August 2018)*.

42. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. *Nat. Genet.* **39**, 126 (2007).

43. Cheeseman, I. H. *et al.* Gene copy number variation throughout the Plasmodium falciparum genome. *BMC Genomics* **10**, 353 (2009).

44. Dharia, N. V *et al.* Use of high-density tiling microarrays to identify mutations globally and elucidate mechanisms of drug resistance in Plasmodium falciparum. *Genome Biol.* **10**, R21 (2009).

45. Volkman, S. K., Neafsey, D. E., Schaffner, S. F., Park, D. J. & Wirth, D. F. Harnessing genomics and genome biology to understand malaria biology. *Nat. Rev. Genet.* **13**, 315–328 (2012).

46. Lourens, C. *et al.* Implementation of a reference standard and proficiency testing

programme by the World Wide Antimalarial Resistance Network (WWARN). *Malar. J.* **9**, 375 (2010).

47. Aurrecoechea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–D543 (2009).

48. Talundzic, E. *et al.* A next-generation sequencing and bioinformatics protocol for Malaria drug Resistance marker Surveillance (MaRS). *Antimicrob. Agents Chemother.* **62**, AAC--02474 (2018).

49. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–85 (1998).

50. Needleman, S. B. & Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

51. Smite, T. F. & Waterman, M. S. *Identification of Common Molecular Subsequences*. *J. Mol. Biol* **147**, (1981).

52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).

54. Hansen, N. F., Gartner, J. J., Mei, L., Samuels, Y. & Mullikin, J. C. Shimmer:

detection of genetic alterations in tumors using next-generation sequence data. *Bioinformatics* **29**, 1498–1503 (2013).

55. Hurgobin, B. Short read alignment using SOAP2. in *Methods in Molecular Biology* **1374**, 241–252 (2016).

56. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).

57. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **12**, 555–560 (2019).

58. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

59. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

60. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. 1–9 (2012). doi:arXiv:1207.3907 [q-bio.GN]

61. Bush, S. J. *et al.* Genomic diversity affects the accuracy of bacterial SNP calling pipelines. *bioRxiv* 653774 (2019). doi:10.1101/653774

62. Zook, J. M. *et al.* Data Descriptor: Extensive sequencing of seven human genomes to characterize benchmark reference materials. (2016). doi:10.1038/sdata.2016.25

63. Cornish, A. & Guda, C. A Comparison of Variant Calling Pipelines Using Genome

in a Bottle as a Reference. *Biomed Res. Int.* **2015**, 1–11 (2015).

64. Chiara, M. *et al.* CoVaCS: a consensus variant calling system. *BMC Genomics* **19**, 120 (2018).

65. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

66. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmarking SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

67. Lee, I., Ouk Kim, Y., Park, S.-C. & Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* **66**, 1100–1103 (2016).

68. Yoon, S.-H., Ha, S., Lim, J., Kwon, S. & Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* **110**, 1281–1286 (2017).

69. Goris, J. *et al.* DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. doi:10.1099/ijs.0.64483-0

70. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

71. Kurtz, S. *et al. Versatile and open software for comparing large genomes*. **5**, (2004).

72. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

73. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59 (2015).

74. Escandón, P. *et al.* Molecular Epidemiology of Candida auris in Colombia Reveals a Highly Related, Countrywide Colonization With Regional Patterns in Amphotericin B Resistance. *Clin. Infect. Dis.* **68**, 15–21 (2018).

75. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

76. Huang, C. K-mer based data structures and heuristics for microbes and cancer. (2018).

77. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

78. Jain, C., Dilthey, A., Koren, S., Aluru, S. & Phillippy, A. M. A fast approximate algorithm for mapping long reads to large reference databases. in *International Conference on Research in Computational Molecular Biology* 66–81 (Springer, Cham, 2017). doi:10.1007/978-3-319-56970-3_5

79. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A Fast Adaptive Algorithm for Computing Whole-Genome Homology Maps. *bioRxiv* 259986 (2018). doi:10.1101/259986

80. Otto, T. D. *et al.* Genomes of all known members of a Plasmodium subgenus reveal

paths to virulent human malaria. *Nat. Microbiol.* **3**, 687–697 (2018).

81. Rutledge, G. G. *et al.* Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. *Nature* **542**, 101–104 (2017).

82. Runtuwene, L. R. *et al.* Nanopore sequencing of drug-resistance-associated genes in malaria parasites, Plasmodium falciparum. *Sci. Rep.* **8**, 8286 (2018).

83. The CRyPTIC Consortium and the 100, 000 Genomes Project, Walker, D., Phil, A. S., Walker, T. E. A. & Peto, D. P. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. *N. Engl. J. Med.* **379**, 1403–1415 (2018).

84. Walker, T. M. *et al.* Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

85. Coll, F. *et al.* Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **50**, 307–316 (2018).

86. Miotto, O. *et al.* Genetic architecture of artemisinin-resistant Plasmodium falciparum. *Nat. Genet.* **47**, 226–234 (2015).

87. Loy, D. E. *et al.* Evolutionary history of human Plasmodium vivax revealed by genome-wide analyses of related ape parasites. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E8450–E8459 (2018).

88. Audano, P. & Vannberg, F. KAnalyze : A Fast Versatile Pipelined K-merizer Supplementary Information. 1–8

89. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

90. Homer, N. DWGSIM: Whole Genome Simulator for Next-Generation Sequencing. *GitHub Repos.* (2010).

91. Pritchard, L. pyani: Python module for average nucleotide identity analyses. (2019).

92. PACE. {P}artnership for an {A}dvanced {C}omputing {E}nvironment ({PACE}). (2017).

93. Audano, P. & Vannberg, F. KAnalyze: A fast versatile pipelined K-mer toolkit. *Bioinformatics* **30**, 2070–2072 (2014).

94. Zhu, S. J., Almagro-Garcia, J. & McVean, G. Deconvolution of multiple infections in Plasmodium falciparum from high throughput sequencing data. *Bioinformatics* **34**, 9–15 (2018).

95. Galinsky, K. *et al.* COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malar. J.* **14**, 4 (2015).

96. Brouwer, R. W. W. *et al.* NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics* **28**, 284–285 (2012).

97. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.1--33 (2013).

98. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts199

99. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54--D56 (2011).

100. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner*. *Joint Genome Institute, department of energy* **13**, (2014).

101. Zaharia, M. *et al.* Faster and More Accurate Sequence Alignment with SNAP. (2011).

102. Grüning, B. *et al.* Practical Computational Reproducibility in the Life Sciences. *Cell Syst.* **6**, 631–635 (2018).

103. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).

VITA

Shashidhar Ravishankar was born in Secunderabad, India and grew up in Bangalore, India. He received a Bachelor of Engineering in Biotechnology, at PES Institute of Engineering in Bangalore India in 2013. After graduation, he worked at InterpretOmics India Pvt. Ltd for one year. He left InterpretOmics in 2014 to pursue a graduate degree in Bioinformatics at Georgia Institute of Technology, Atlanta.