A suite of computational tools to interrogate sequence data with local haplotype analysis within complex Plasmodium infections and other microbial mixtures

A suite of computational tools to interrogate sequence data with local haplotype analysis within complex *Plasmodium* infections and other microbial mixtures

A Dissertation Presented

By

NICHOLAS JOHN HATHAWAY

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

MARCH 19, 2018

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
M.D., PH.D. PROGRAM

A suite of computational tools to interrogate sequence data with local haplotype analysis within complex *Plasmodium* infections and other microbial mixtures

A Dissertation Presented

By

NICHOLAS JOHN HATHAWAY

This work was undertaken in the Graduate School of Biomedical Sciences

Bioinformatics and Computational Biology

Under the mentorship of

Jeffrey Bailey, MD, PhD, Thesis Advisor

Elinor Karlsson, PhD, Member of Committee

Manuel Garber, PhD, Member of Committee

Daniel Neafsey. PhD, External Member of Committee

Zhiping Weng, PhD, Chair of Committee

Mary Ellen Lane, PhD,

Dean of the Graduate School of Biomedical Sciences

MARCH 19, 2018

Dedication

This work is dedicated to my friends and family who have supported me emotionally during my training and to Chipotle which has supported me physically.

Acknowledgements

The last five years would not have been possible without a great many people and I am sure I will miss acknowledging many. I would like to thank Christian Parobek, Robin Miller, and Sujata Balasubramanian for being my first beta testers when I was still very new to programing and had to suffer through a lot of bugs. I would like to thank Michael Purcaro, who's constant support and advice has help shaped me into the programmer I am today. I would also like to thank Arjan van der Velde who has put up with all my miscellaneous questions with anything computer science or network related. I would also like to give appreciation to all my collaborators for making research both intellectually stimulating and also very entertaining; all past and present members of IDEEL at UNC including but exclusive to Jonathan Juliano, Steven Meshnick, Jessica Lin, Jonathan Parr, Christian Parobek, Jaymin Patel, Sujata Balasubramanian, Nicholas Brazeau, Andreea Waltman, Molly Deutsch-Feldman, from Imperial College London, Robert Verity and Oliver Watson, and from UCSF, Bryan Greenhouse and Sofonias Tessema among many others. I would also like to acknowledge the great amount of support given by Barbara Bucciaglia, Heidi Beberman, and Christine Tonevski, who have helped to make my day to day so much easier.

I would also, of course, like to thank my own lab for their support; Özkan Aydemir, Yasin Kaymaz, Patrick Marsh, Jennifer Moon, and Mercedeh Javanbakht Movassagh among others and to thank my mentor Jeffrey Bailey for taking a chance on me when I had little experience and allowed me to join his lab making the work I present here possible. I would also like to thank my great friend Tim Nicastro, who has help me edit my thesis into something human readable and for being a constant source of support in my life.

Abstract

The rapid development of DNA sequencing technologies has opened up new avenues of research, including the investigation of population structure within infectious diseases (both within patient and between populations). In order to take advantage of these advances in technologies and the generation of new types of data, novel bioinformatics tools are needed that won't succumb to artifacts introduced by the data generation, and thus provide accurate and precise results. To achieve this goal I have create several tools.

First, SeekDeep, a pipeline for analyzing targeted amplicon sequencing datasets from various technologies, is able to achieve 1-base resolution even at low frequencies and read depths allowing for accurate comparison between samples and the detection of important SNPs. Next, PathWeaver, a local haplotype assembler designed for complex infections and highly variable genomic regions with poor reference mapping. PathWeaver is able to create highly accurate haplotypes without generating chimeric assemblies. PathWeaver was used on the key protein in pregnancy associated malaria *Plasmodium falciparum* VAR2CSA which revealed population sub-structuring within the key binding domain of the protein observed to be present globally along with confirming copy number variation. Finally, the program Carmen is able to utilize PathWeaver to augment the results from targeted amplicon approaches by reporting where and when local haplotypes have been found previously.

These rigorously tested tools allow the analysis of local haplotype data from various technologies and approaches to provide accurate, precise and easily accessible results.

List of Tables

List of Figures

Chapter 1: Introduction

The development of high-throughput sequencing has helped to advance the field of molecular genetics, with the ability to generate data rapidly outstripping our ability to analyze it. With these advances in technologies constantly increasing the amount that can be sequenced and advances making sequencing cheaper, more and more research fields are performing sequencing analyses. This has necessitated the need for novel bioinformatics tools that are able to analyze a large amount of data at once and that can be used by researchers of many different skill levels. This means the output of results needs to be in a form that is most readily consumable by other analysis pipelines and researchers in order to ensure the most efficient workflows.

The availability of high-throughput sequencing has been a great asset to the study of microbial populations, like the analysis of bacterial communities (Taft et al. 2015), viral quasispecies analysis (Beerenwinkel et al. 2012), and malaria infections (Lin et al. 2015; Mideo et al. 2016), among other infectious diseases. The analysis of microbial populations differs from sequencing approaches carried out on a human because rather than sequencing a single genome from one individual, there exists a population of individuals (microbes) with several genomes present. These populations are clonal populations which means that many of the individuals have the same exact genome and the population can either be monoclonal, meaning all individuals have the same exact genome, or polyclonal, a mix of genomes with varying degrees of differences with some individuals sharing the same exact genome. These polyclonal populations are often referred to as “complex” mixtures.

The study of the *Plasmodium* species, the causative agent of malaria, has benefited greatly from the advances in high-throughput sequencing. *Plasmodium* is a protozoan parasite transmitted by female *Anopheles* mosquitoes and there are five known species to infect humans; *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae*, and *P. knowlesi,* with *P. falciparum* being credited with the majority of the fatalities. Plasmodium was estimated to infect 216 million people and caused 445,000 deaths in 2016 (WHO 2017). Immunity to *Plasmodium* is not sterilizing allowing for repeated infections and people living in endemic areas eventually stop experiencing symptomatic infections by various processes but mostly thought to be achieved by building up a repertoire of antibodies to keep parasitemia low (A. Barry and Hansen 2016). The life cycle of *Plasmodium* is complex; it is diploid within its mosquito vector and haploid while it infects its human host. It undergoes several stages within the human host including a liver stage and a blood stage where it spends the majority of its time inside of infected erythrocytes and the species *P. vivax* and *P. ovale* can lay dormant in the liver.

Plasmodium is a long time enemy of humans and has been credited with being one of the most powerful recent forces for causing genetic changes within the human genome (Evans and Wellems 2002; McManus et al. 2017). This genetics arms race between humans and *Plasmodium* has lead to the creation of extreme diversity within *Plasmodium*, driven in part by balancing selection and directional selection (Weedall and Conway 2010).

Balancing selection is a phenomenon that selects for diversity, especially in immune epitopes; the more diverse an epitope, the more likely the parasite is able to survive and reinfect a host with a previous infection especially if cross-strain reactive antibodies are not able to be formed. This is especially true for infectious agents that don’t induce lasting immunity and lead to individuals being infected multiple times which leads to a strain’s

frequency being inversely correlated with its survivability (Lipsitch and O'Hagan 2007). The presence of balancing selection is important for vaccine candidate consideration as epitopes that exhibit balancing selection will likely not make good vaccine candidates because the parasite will have the necessary diversity to avoid the antibodies induced by the vaccine. For this reason many vaccines for *Plasmodium* have failed to induce completely protective antibodies (Offeddu et al. 2012; Ouattara et al. 2013) with the most effective vaccine being only 35% protective (Neafsey et al. 2015). It has been shown that even a single base difference between strains can prevent cross reactivity of antibodies (Sedegah et al. 2016) and therefore it is essential to be able to achieve single base resolution for the study of vaccine candidate regions within *Plasmodium* for both informing the development of a potential vaccine and for the monitoring of current vaccines.

While balancing selection maintains diversity and prevents allelic fixation, directional selection promotes allele fixation. Common examples of directional selection include mutations that help parasites to adapt to a new host and mutations that induce drug resistance. These pressures will cause a single mutation to fix rapidly within a population shortly after occuring. *P. falciparum* has a long history of developing drug resistance and often resistance can form with just a single base change like the K76T mutation in the chloroquine resistance transporter (CRT) (Lakshmanan et al. 2005) among several other examples (Basco et al. 1995; Nagesha et al. 2001; Nwakanma et al. 2014). As these mutations are not selected for until the pressure of a specific drugs is added, mutations can stay at low frequencies in populations and it is important for drug resistance monitoring purposes to be able to detect these single base differences even at these low frequencies to help predict the possibility of drug treatment failure (Ngondi et al. 2017).

One method for analyzing *Plasmodium* diversity is to focus analysis on specific genomic regions of interest by designing PCR primers to amplify and sequence only these regions, an approach called targeted amplicon analysis (Lerch et al. 2017; Hathaway et al. 2017). Depending on the specific research goals, the targeted regions could be genes responsible for drug resistance (Ngondi et al. 2017), vaccine candidates (Bailey et al. 2012; Neafsey et al. 2015; Mideo et al. 2016), highly diverse surface antigens targeted by the immune system which make for good biomarkers to trace strains (R. H. Miller et al. 2017; Patel et al. 2017; Verity et al. 2018; Lin et al. 2015) or regions that are associated with disease mechanisms (Patel et al. 2017; Waltmann et al. 2018). Complex *Plasmodium* infections, being infected by more than one strain at a time, are a common occurrence in endemic areas (Juliano et al. 2010; Arez et al. 2003); as a result, in some research approaches, samples from multiple patients are pooled together, a common technique for studies on drug resistance mutations monitoring (Taylor et al. 2010; Ngondi et al. 2017). Therefore it is important that any analysis pipeline for targeted amplicon sequencing to be able to detect differences between strains within a mixture with single base resolution and at low abundances.

In order to accomplish accurate targeted amplicon analysis from high-throughput sequencing, the various errors produced in the data generation need to be corrected without over-correcting by removing real biological variation or by under-correcting by reporting error as true biological variation. The major sources of error are in the PCR amplification step and the sequencing step. The errors produced by PCR include the creation of single base substitutions which can result at appreciable frequencies depending on the amount of input DNA and the PCR cycle within which the error occurred (e.g. errors in early cycles of PCR,

especially for low input DNA amounts, will prograte into the next cycles). The errors produced in the sequencing step will be dependent upon the specific technology used.

There now exist many different sequencing technologies which all vary in sequencing methodologies and the number and length of sequences they can produce among many other aspects (Quail et al. 2012). While technologies share some similarities in the data they produce (e.g. most technologies will supply “per base quality scores,” a score representing how likely a base call is a sequencing error or not), each comes with its own set of specific biases that might have to be handled differently. For example, 454 and Ion Torrent create a large number of indels in homopolymers (long stretches of the same nucleotide base) due to the similarity in sequencing methodology they use; however, they also both compute their per base quality scores in very different ways, and thus special care is needed when utilizing the quality scores they report (Brazeau et al. 2016). Also, while Illumina, 454, and Ion Torrent tend to have decreasing quality scores and increased error as they move along a read, technologies like PacBio have high error rates that have no correlation with position in the read. For this reason, care must be taken when trying to apply one computational tool built for a specific technology to another technology; certain assumptions based on one technology's characteristics could be invalid on another technology and lead to artifacts.

The majority of analyses with targeted amplicon approaches have been conducted determining various microbiomes by sequencing various variable regions of the 16S ribosomal subunit (NIH HMP Working Group et al. 2009) and this has greatly influenced the tools that are available (Caporaso et al. 2010; Edgar 2013). The 16S subunit is analyzed because it is shared by all bacterial species but, because of its extreme biological importance, it is slowly evolving and doesn’t differentiate bacteria at the species level (Woo

et al. 2008; Janda and Abbott 2007). In fact, the most closely related 16S sequence are often different 16S copies within one bacterial species genome (Kembel et al. 2012). For this reason the majority of 16S targeted amplicon sequencing is conducted by clustering sequences at an operational taxonomic unit (OTU) which involves clustering sequences that are similar up to a certain percent identity, commonly 97%. Clustering at 97% identity would serve to correct for any errors that arise in data generation but would also collapse many real biological differences which could be up to 9 differences for a region of 300 bases in length. Clustering at this level is not adequate for studying *Plasmodium* and a greater resolution is needed that would still correct for any errors. While some recent developments have greatly increased the resolution capable, these approaches either are unable to detect single base differences, like Swarm (Mahé et al. 2014), or fail to detect single base differences at low read depths or low frequencies like the programs UNOISE2 (Edgar 2016) and DADA2 (Callahan et al. 2016).

Shotgun whole genome sequencing is another popular technique for analyzing *Plasmodium* infections. This process involves generating reads from the entire genome of interest, rather than just a targeted locus like in targeted amplicon sequencing. While this generates a large amount of data across the whole genome, reads all start and end in random locations and often require mapping to a reference genome to be analyzed. These mapped reads are normally run through a traditional variant-calling pipeline for calling single polymorphisms (SNPs) and short insertions and deletions (INDELs). While this process works for stable regions of the genome, it might fail to call variants in regions so diverse that mapping to a reference is not possible. For this reason, highly diverse regions of the *Plasmodium* genome are often masked from this type of analysis even though these regions encode key virulence factors for *Plasmodium*. There is a growing large collection of publicly

available field samples that have been shotgun whole genome sequenced has been generated but remain unanalyzed for these key virulence factors.

One of the major contributing factors to *P. falciparum* causing more fatal clinical outcomes is due to its multigene family called *var* genes that encode for the protein erythrocyte membrane protein 1 (PfEMP1) (Smith et al. 2000; Gardner et al. 2002). This highly diverse family encodes a protein that contains several domains capable of binding to various endothelial cell surfaces. The protein gets transported to the surface of *P. falciparum* infected erythrocytes and causes the erythrocytes to adhere to the walls of blood vessels. This helps the parasite survive by avoiding traveling through the spleen which is a major mechanism for clearing infected erythrocytes (Rowe et al. 2009). The adhesion of infected erythrocytes to blood vessel walls can lead to the destruction of microvasculature and depending on the location of the microvascular can lead to various clinical outcomes (e.g. destruction of microvasculature in the brain can lead to stroke) (Rowe et al. 2009). There are approximately 60 *var* genes each capable of binding to various targets and many of the genes undergoing recombinations between different *var* genes (Rask et al. 2010; Gardner et al. 2002). The *var* genes have a complex transcription regulation and only one *var* gene is expressed at a time (Duffy et al. 2017; Dimonte et al. 2016).

One *var* gene of interest that appears to have become isolated from the other *var* genes and only undergoes recombination with itself is VAR2CSA which binds to chondroitin sulfate (CSA), a protein only found on placental tissue (Salanti et al. 2003a). Therefore, VAR2CSA expressing *P. falciparum* parasites that infect a pregnant woman can cause the destruction of placental tissue and poor birth outcomes (Salanti et al. 2003a). Naturally-acquired antibodies to VAR2CSA have been shown to be protective during pregnancy (Rogerson et al. 2007; Ataíde, Mayor, and Rogerson 2014). Efforts to develop a

VAR2CSA vaccine are underway (Fried and Duffy 2015; Tuikue-Ndam and Deloron 2015). However, their efficacy may be hampered by the genetic and geographical variation in the protein. Previous studies on VAR2CSA diversity have been limited to less than 30 full length genes and the full global diversity has yet to be fully revealed.

The *var2csa* gene is approximately 8 kilobases (kb) long and key binding domains with the protein are 1.8kb. There are few conserved regions, due to its high diversity, to target for PCR and the gene's long length make a targeted approach unfeasible. This high diversity also prevents traditional variant calling with reference based mapping approaches as the sequence differences prevent approximately 20% of the reads from mapping to the reference. Due to this poor mapping which is common to all *var* genes, previous attempts have been made to use assembly programs, like SPAdes (Bankevich et al. 2012), to construct *var* gene sequences (Jespersen et al. 2016; Lennartz et al. 2017). However, these previous attempts have not been extensively validated and it has been observed that assemblies done on polyclonal mixtures can lead to erroneous chimeric assemblies where sequences from different genomes are combined into one sequence. The assembly programs being used were designed to do assembly of only one genome and therefore do not handle the presence of multiple genomes well.

In this thesis, I present several novel computational tools I have created to analyze the high diversity found within microbial mixtures with a focus on *Plasmodium*. Chapter II describes the program SeekDeep for analyzing targeted amplicon sequencing to achieve single base resolution even at low frequencies. Chapter III extends the use of SeekDeep to be used on longer amplicon targets created by the PacBio technology. Chapter IV presents PathWeaver, a program designed to recruit *var2csa* sequences from shotgun whole genome sequencing datasets to assemble highly diverse regions of the genome, which I apply to the

var2csa gene to fully elucidate the global diversity to help further vaccine development. Chapter V further leverages the PathWeaver program to utilize the wealth of publicly available data to augment targeted amplicon analysis by reporting on where and when haplotypes have been found previously.

Chapter II: SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing

Preface

The following research chapter was adapted from “SeekDeep: Single-Base Resolution de Novo Clustering for Amplicon Deep Sequencing.” Hathaway, Nicholas J., Christian M. Parobek, Jonathan J. Juliano, and Jeffrey A. Bailey. 2017 *Nucleic Acids Research*, November. https://doi.org/10.1093/nar/gkx1201. (Hathaway et al. 2017).

ABSTRACT

PCR amplicon deep sequencing continues to transform the investigation of genetic diversity in viral, bacterial, and eukaryotic populations. In eukaryotic populations such as *Plasmodium falciparum* infections, it is important to discriminate sequences differing by a single nucleotide polymorphism. In bacterial populations, single-base resolution can provide improved resolution towards species and strains. Here we introduce the SeekDeep suite built around the qluster algorithm, which is capable of accurately building *de novo* clusters representing true, biological local haplotypes differing by just a single base. It outperforms current software, particularly at low frequencies and at low input read depths, whether resolving single-base differences or traditional OTUs. SeekDeep is open source and works with all major sequencing technologies, making it broadly useful in a wide variety of

applications of amplicon deep sequencing to extract accurate and maximal biologic information.

INTRODUCTION

The development of targeted next-generation sequencing technologies has dramatically expanded research into population-level genetic diversity, from the study of bacterial communities (Taft et al. 2015), intrahost variation in infections, such as HIV and malaria (Beerenwinkel et al. 2012; Lin et al. 2015; Mideo et al. 2016), to heterogeneity in cancer tumors (Dawson et al. 2013). In general, targeted amplicon deep sequencing utilizes areas of conserved sequence for amplification primer placement, surrounding a region of interest containing known mutations or high sequence variability. Thousands to millions of product molecules from the amplification are then individually sequenced using current massively parallel techniques. To date, experimental and computational techniques for deep sequencing have been driven largely by microbiome 16S and targeted viral sequencing where single-base resolution is not a necessity (Quince et al. 2011; Beerenwinkel et al. 2012; Prabhakaran et al. 2010). While initial microbiome work has focused on genus-level resolution of 97% sequence identity, there is greater interest in maximizing species and strain information in bacterial and viral populations (Benítez-Páez, Portune, and Sanz 2016; Beerenwinkel and Zagordi 2011). In eukaryotic populations, such as malaria strains, and for mutation detection, differentiation at the single-nucleotide level resolution is a necessity (Lin et al. 2015; Mideo et al. 2016).

The central bioinformatic challenge of all targeted deep sequencing is to accurately resolve the true biologic differences that are obscured by the numerous errors introduced during PCR amplification and sequencing. PCR errors include substitutions, insertions and

deletions, as well as chimeras formed by incomplete extension and subsequent re-priming on a highly-similar (but non-identical) template (**Figure 2.1**). Sequencing error types and frequencies tend to be platform specific, and are related to either the sequencing polymerase or detection technology. For instance, pyrosequencing-based technologies generate numerous insertion-deletion (indel) errors, particularly in homopolymers, since these technologies estimate the number of a particular nucleotide in succession based on the cumulative fluorescent (454) or ion (Ion Torrent) signal. On the other hand, Illumina technology mainly misidentifies individual nucleotides, thus producing base-substitution errors (Lysholm, Andersson, and Persson 2011; Huang et al. 2012).

Numerous computational solutions have been developed to correct for these errors (Zhbannikov and Foster 2015), including minimum entropy decomposition (MED (Murat Eren et al. 2014)), homopolymer runs correction (Acacia (Bragg et al. 2012)), clustering based on consistency of inferred error models (DADA2 (Callahan et al. 2016)), operational taxonomic unit (OTU) clustering (UPARSE (Edgar 2013)), k-mer correcting (KEC (Skums et al. 2012)), and many others (Zagordi et al. 2011; Yang, Chockalingam, and Aluru 2013). All of these methods have advantages and disadvantages vis-a-vis speed, sensitivity, specificity, flexibility, range of sequencing technologies, and types of errors corrected. In general, the latest methods aim for greater resolution to allow better definition of microbial populations. The ultimate goal is discriminating sequences differing by a single base, which is the quantum level of evolutionary change. Such resolution will allow more detailed assessment of bacterial, viral, and eukaryotic microbial populations particularly with longer amplicons. Consistent single-base resolution is a particular necessity for studies of eukaryotic intra-species populations and for mutation detection. For example, in malaria research, the sequence of a single amplicon is frequently used to define strains within an

infected individual, and these sequences often differ by only a single base, representative of a SNP within the larger parasite population. In microbiome studies, single-base resolution of 16S amplicon clustering extracts maximal information for downstream analyses. Thus, we sought to develop new algorithms that could consistently differentiate single-base differences in a wide variety of conditions and applications including improved accuracy and sensitivity of traditional operational taxonomic units (OTUs).

Here we present SeekDeep, an open-source software suite for *de novo* (i.e reference free) analysis of amplicons that is fast, sensitive, customizable, and is able to resolve sequences differing by only a single base, even at low frequencies. At the center of SeekDeep is the algorithm qluster (for quality clustering) that improves the correction of PCR and sequencing errors in multiple key ways including base quality values and k-mer frequencies. SeekDeep also provides a growing set of pre- and post-processing tools, including an embedded web server to dynamically view results and ancillary data - particularly useful when working with large datasets and numerous samples, a scenario which has become common with targeted amplicon studies (Lin et al. 2015; Mideo et al. 2016).

We compared SeekDeep to other recent best-in-class programs, DADA2 (Callahan et al. 2016), MED (Murat Eren et al. 2014) and UNOISE in USEARCH (preprint https://doi.org/10.1101/081257), which also aim for single-base resolution. All programs aim to determine the local PCR amplicon haplotypes, herein referred to simply as haplotypes for brevity, that represent the specific sequences (linked variation from the same chromosome) found in the biologic material prior to amplification. We also compared OTU based clustering to commonly used programs USEARCH (aka UCLUST/UPARSE) (Edgar 2013) and to Swarm (Mahé et al. 2014) which cannot resolve at the single-base level. We focused our

comparisons on programs that could work with both Illumina and 454/Ion Torrent sequence and did not compare to programs that only correct 454 and Ion Torrent pyrosequencing errors like AmpliconNoise (Quince et al. 2011), Acacia (Bragg et al. 2012), and HECTOR (Wirawan et al. 2014) as we are interested in tools that are broadly applicable in the field.

To ascertain the performance of these programs, we compared results of *in silico* simulated datasets and *in vitro* mixtures of isolated DNA representing mock infections of both *Plasmodium falciparum* and bacterial communities. The simulations focused on the quantitative accuracy of discerning minor (low-abundance) haplotypes in terms of how much they differ (1 to 13 bp equating to 99.6% to 95.6% similarity) from a major (high-abundance) haplotype and how much they differ from another minor haplotype unrelated to all other haplotypes.

RESULTS

Simulation Studies

First we compared the performance of SeekDeep to the other programs on the two types of simulated mixtures: mixtures where minor haplotypes are closely related to a major haplotype which was at a much greater abundance (**Figure 2.2a-b**), and mixtures where a minor haplotype was closely related to another minor haplotype at the same abundance (**Figure 2.2c-d**). For all simulations, SeekDeep matched or outperformed MED, DADA2, and UNOISE in recovery of all haplotypes, especially one-off haplotypes (**Figure 2.3**). SeekDeep showed improved haplotype recovery compared to other methods, which was accentuated as read depth, divergence and abundance of haplotypes decreased (**Figures 2.3, 2.4-2.6**). Together these factors combined to show marked differences in haplotype recovery for

low-abundance haplotypes differing by a single-base (i.e. one off from the closest sequence) assessed with low numbers of reads (**Figure 2.3**). SeekDeep was also better able to estimate the expected abundance of haplotypes, demonstrated by a lower root mean squared error (RMSE) (**Figures S12-13**) compared to all programs. The MED algorithm appears to have trouble as a haplotype's abundance increases, which could be due to the fact that it was developed specifically for microbiome data where the abundance of each haplotype usually does not exceed more than 10%. Though SeekDeep creates more false haplotypes than DADA2 and UNOISE, the abundance of the false haplotypes is generally much lower than 0.1% while DADA2, MED, and UNOISE were shown, especially for 454, to create false haplotypes greater than 1%, with most falling between 0.1% and 1% (**Figure 2.9**). While DADA2 minimizes the number of false haplotypes (**Figure 2.9**), it also loses sensitivity particularly with lower read depth input (**Figures 2.3, 2.5**). Overall, SeekDeep shows greater consistency at lower thresholds providing unbiased detection in the face of variable haplotype abundance and input read depths.

In vitro Control Mixtures

Plasmodium 454 and Ion Torrent pyrosequencing

Next we evaluated the performance of haplotype detection for *P. falciparum* lab strains for *TRAP*, *AMA1*, and *CSP* genes on both 454 and Ion Torrent by creating mock mixtures in the lab that were PCR amplified and then sequenced. This provides important insight into factors that may not be captured in the simulated sequence. For these *in vitro* mixtures, both SeekDeep and MED were able to achieve 100% haplotype recovery across all samples while UNOISE had 92% and DADA2 had 83% haplotype recovery (**Figure 2.10a**). Missed haplotypes usually represented the collapse of low-abundance highly similar

haplotypes. In part, it also appeared that haplotype recovery for UNOISE and DADA2 were hampered by indel errors especially in homopolymers which are difficult to overcome in Ion Torrent and 454 data. This is a known issue as UNOISE's website states that UNOISE does not work well on Ion Torrent and 454 data (http://www.drive5.com/usearch/manual/faq_unoise_not_illumina.html). All programs had appreciable false haplotypes (**Figure 2.10c**), and, while DADA2 had the lowest number of false haplotypes, when they did occur they often had appreciable frequencies even exceeding 10%. Only SeekDeep limited the occurrence of false haplotypes to low abundances (<=0.5%). Replicates again aided all programs but dramatically reduced the number of false haplotypes for SeekDeep. SeekDeep again showed the most accurate abundance estimates (**Figure 2.10b**). Notably MED, while demonstrating 100% haplotype recovery, consistently underestimated abundances due to the numerous false haplotypes at appreciable frequencies (**Figure 2.10c**).

Plasmodium Illumina MiSeq

We also evaluated a mock mixture of *P. falciparum* across 23 loci that represent important markers of drug resistant or regions of diverse variation. These amplicons were PCR amplified and sequenced on Illumina MiSeq 2x250 paired end. SeekDeep and MED were able to achieve 100% haplotype recovery of all 23 targets while DADA2 and UNOISE both failed to detect nine out of the 88 total haplotypes. Five haplotypes were missed in common by both programs (**Figures 2.11 and 2.12**). The haplotypes that UNOISE and DADA2 failed to detect where either related to another haplotype by a single nucleotide or 1 large indel (~10 nucleotides) and ranged in abundance from 4 - 20%. SeekDeep demonstrated a minimal number of false haplotypes on par with UNOISE (**Figure 2.11c**).

Unlike UNOISE and the other programs which report false haplotypes at abundances that can exceed 10%, SeekDeep's false haplotypes were all less than 0.8% abundance. Again SeekDeep showed the highest accuracy in terms of predicting the abundance (**Figure 2.11b**).

Mock Microbiome

We also tested SeekDeep on a mock microbiome dataset previously described in Salipante *et al*. 2014 (Salipante et al. 2014), which had been amplified and sequenced in triplicate on the Illumina platform. It contained 47 distinct 16S copies (**Table 2.1)**. MED and SeekDeep were able to recover 100% of all expected haplotypes in all datasets, while DADA2 missed one haplotype. For all three replicates of this dataset, DADA2 missed the *L. monocytogenes*.2 haplotype, which had an expected abundance of 0.8% and is one nucleotide different from the *L. monocytogenes*.5 haplotype which had an expected abundance of 1.5%. UNOISE also missed *L. monocytogenes*.2 in one replicate and in all three replicates missed *B. vulgatus.3* (0.035%), *B. cereus.4* (0.33%), and *B. cereus.1* (0.36%), haplotypes, which all differ by one nucleotide from another haplotype.

Downsampled Mock Microbiome

Because the mock microbiome dataset previously described in Salipante *et al*. 2014 (Salipante et al. 2014) was sequenced to a great depth (>600,000 reads), we randomly downsampled the dataset to lower read depths (2,000-20,000) to test detection at levels of sequencing more commonly employed in experiments. For the downsampled mock microbiome dataset from Salipante *et al*. 2014 (Salipante et al. 2014), SeekDeep outperformed DADA2, MED, and UNOISE in haplotype recovery of the twenty-three one-off haplotypes (out of forty-seven total haplotypes in the dataset). The highest relative

abundance of missed one-off haplotypes was approximately 3% for DADA2 and MED, 2% for UNOISE, but only 0.25% for SeekDeep (**Figure 2.13**). Again, SeekDeep does well using fewer input reads in estimating the expected abundance of the known haplotypes with a lower RMSE (**Figure 2.14**).

Viral strain mixtures

To further ensure that SeekDeep works across a breadth of experiments and organisms, we examined control mixtures of viral strains. All programs performed well with respect to recall for both the EBV (**Figure 2.15)** and HIV (**Figure 2.16)** mixtures, which was not unexpected as these mixed strain haplotypes all differed by more than a single base. Importantly, SeekDeep's specificity compared well. All other programs other than SeekDeep created false haplotypes above 1% in the EBV dataset with the highest for each program being 2% for UNOISE, 17% for MED, 22% for DADA2, 19% for ShoRAH and 0.65% for SeekDeep which was mitigate but not completely removed with replicates (1% for UNOISE, 16% for MED,14% for DADA2, 16% for ShoRAH and 0.46% for SeekDeep). Programs performed better on the HIV dataset and though SeekDeep had a large number of false haplotypes all of them fell below the recommended cut off of 0.5% with the highest being at 0.35%. This high amount of apparent false haplotypes at low frequency was probably representative of both increased biologic variation due to HIV replication by error-prone reverse transcriptase as well as the elevated 65 rounds of PCR amplification prior to sequencing.

Chimera detection

For these *in vitro* control mixtures, chimera formation and abundance was highly variable depending upon the experiment. The Illumina *P. falciparum* dataset only demonstrated 3 chimeras across all 28 amplicions. The IonTorrent controls demonstrated significant numbers of low abundance chimeras. Across the 7 samples there were a total of 186 false haplotypes of which 83% (155) were chimeras. These IonTorrent false haplotypes generally showed higher abundances relative to other false haplotypes and were highly-reproducible abundances across replicates (R^2=0.81-0.99; **Figure 2.17**). The differences in chimera formation between datasets most likely originates from differences in the amount of input template and PCR conditions as well as potentially the library preparation which involves PCR. The mock microbiome showed minimal chimera formation likely due to the decreased sequence relatedness and greater amounts of starting template. Overall, the variability in chimera occurrence rates along with their high-degree of reproducibility within replicates emphasizes the need to carefully consider the experimental conditions and the utilization of experimental controls to determine the need and optimal settings for chimera detection.

Traditional Microbiome OTU Analysis

In addition to providing single-base resolution between sequences, SeekDeep was designed to also allow users to define the needed level of resolution by setting either the number of bases or percent identity to create operational taxonomic units (OTUs). We therefore compared SeekDeep to older commonly used programs offering OTU level resolution that can operate on multiple platforms. In comparison to USEARCH (i.e.

UCLUST), Seekdeep showed both better accuracy and precision clustering at 97% OTUs (**Figure 2.18**). Also, USEARCH at times misconstructs the OTUs, returning a consensus sequence that is not one of the actual input haplotypes (**Figure 2.18b**). SeekDeep routinely returns the major haplotype within an OTU. We also compared to SWARM collapsing on 1-base differences - the most sensitive setting for SWARM (**Figure 2.19**). Again SeekDeep demonstrated better haplotype recovery and fewer false haplotypes. Thus, SeekDeep provides more optimized OTU definition, which again is more robust to varying read depth.

Performance

Algorithm speed can be an important factor in terms of practicality, and SeekDeep compares favorably with other programs. While UNOISE is the fastest algorithm (**Figure 2.20**), this speed comes at a cost (**Figure 2.3**). The proprietary algorithm in UNOISE works by collapsing one-off errors if the ratio of abundance between two sequences is at a certain threshold, which precludes UNOISE from detecting new haplotypes that differ by only one nucleotide from the major haplotype in the population. This aspect can be problematic when screening for cancer mutations or pathogen drug resistance. Also UNOISE recommends not using singlet sequences, decreasing haplotype recovery at lower read depths. This, in part, contributes to its speed (**Figure 2.20**) but decreases haplotype recovery. Apart from UNOISE, SeekDeep is comparable in speed to DADA2 and MED (**Figure 2.20**). In fact, for the mock microbiome data set (Salipante et al. 2014), which had approximately 800,000 for each of three replicates, the run times for the programs were 2hrs and 41 minutes for SeekDeep, 2hrs and 40 minutes for DADA2, and 1hr and 58 minutes for MED on standard hardware as found in a personal computer. Given runtimes are comparable, the built-in general pipelines for sample processing make SeekDeep a potentially less-time consuming

option for the general user looking to process numerous samples and multiple amplicons per sample.

DISCUSSION

With newer sequencing technologies increasing our ability to probe a wide variety of biologic samples, the ability to bioinformatically discern the full extent of sequence diversity, even if only a single-base difference, is key to answering many important questions. Though all programs tested are able to detect one-off haplotypes, SeekDeep is the only one consistently able detect these haplotypes at lower frequencies and at lower input read depths for all technologies (**Figures 2.3**, **2.4-2.6**, **2.13**). SeekDeep performs well across a diverse set of simulations and *in vitro* control data sets and provides a more favorable balance between haplotype recovery and false haplotypes such that missed haplotypes and false calls are limited to the lowest frequencies, usually below 0.25%. In fact, when applying 0.25% as a lower threshold, SeekDeep has near perfect haplotype recovery and precision (**Figure 2.21**). We apply a slightly higher cutoff of 0.5% as the default in processClusters, the final processing step, ensuring high confidence in the called haplotypes. Essentially, SeekDeep provides the ability to confidently detect haplotypes across variable read depths regardless of haplotype abundance, similarity or platform, a feature which is crucial for maximizing experimental information and minimizing biases. Minimizing bias is important for downstream analyses such as time series that generally presume random deviations (Bucci et al. 2016; Friedman and Alm 2012).

SeekDeep showed important differences in terms of haplotype recovery and false haplotype creation compared to other programs. While DADA2 creates a smaller number of false haplotypes than SeekDeep, this comes at the cost of missing low-abundance one-off

haplotypes. Also, when DADA2 does create a false haplotype it is generally at a higher abundance than SeekDeep (**Figures 2.10-2.11, 2.9**). DADA2 did not compare well at lower read depths where haplotype recovery suffered remarkably. Thus, for users of DADA2 it may be important to ensure that all samples have deep read depth to minimize biases. MED has good haplotype recovery but also creates numerous false haplotypes, particularly high-abundance haplotypes in samples with low diversity (**Figures 2.10-2.11, 2.17**). SeekDeep's balance between haplotype recovery and false haplotypes at very low-abundances was by design. For subsequent aggregate or longitudinal analyses across samples, low-level noise in individual samples can often be better controlled across the entire sample set. However, missing haplotypes or false calls at appreciable levels are more difficult to compensate for and can be a source of significant bias.

Importantly SeekDeep is extremely robust to sequence quality or types of sequence variation. SeekDeep directly utilizes the actual base quality of each sequencing read. Thus, it is robust to sequences that are outliers with extremely poor quality. Unlike both MED and DADA2 that require that input sequences be the same exact length, SeekDeep can handle variable length input given it performs optimal global alignments, and thus is adept at analyzing sequences with insertions or deletions. Variable length inputs are very common among Ion Torrent and 454 sequencing data.

Users can further optimize SeekDeep for more advanced applications. It can flexibly cluster based on insert size, allowing for the detection of biologically relevant insertions such as nucleotide triplets consistent with an amino acid change while filtering out homopolymer or smaller indels that are particularly common in some sequencing platforms such as IonTorrent. With SeekDeep, users can set the specific number of each type of alignment differences (indels and/or SNPs) upon which to collapse clusters, enabling concrete tuning

for the specific biologic questions. For instance, this allows a user to collapse haplotypes that differ by one base, two bases, or traditional analyses collapsing to 97% or 99% OTUs and detect more divergent lower abundance variants that may be only represented by a few sequences in a sample.

SeekDeep offers robust and flexible pre- and post-clustering tools and workflows for rapidly preprocessing numerous samples by demultiplexing barcodes, identifying and removing primers, trimming, and cleaning sequence to user specifications. After clustering, the tool set helps evaluate the sequence and perform initial data evaluation with key sample and population statistics. SeekDeep has built-in support for a number of steps including (1) scanning for contamination, which is especially helpful, for example, in *Plasmodium* datasets which can often be contaminated with human DNA due to low relative amount of parasite DNA, (2) built-in support for incorporating replicate comparison, and (3) support for analysis of multiple amplicon targets at once. It also supports chimera detection and removal akin to other programs which should be carefully considered and tuned based on experimental conditions, controls and the biologic question of interest. SeekDeep also provides a dynamically interactive HTML viewer, which makes it easy to explore differences between strains and has support for viewing results on subgroups in large sample sets when given group metadata.

Overall, SeekDeep expands the potential for *de novo* amplicon clustering - particularly given its improved haplotype recovery at lower read depths for haplotypes differing by one base. This is crucial for projects that seek to detect and quantify minority haplotypes that may be represented by a single SNP. Such projects are becoming increasingly common in the oncology and infectious disease fields. For example, when using marker regions to differentiate bacterial strains, or when monitoring for pathogen

drug-resistance mutations, these sequences often only differ from the wild type by a single base (Miotto et al. 2015). Accurately quantifying these low-abundance and genetically similar strains in these cases is key.

In summary, SeekDeep can be widely applied to all forms of amplicon deep sequencing to improve the haplotype recovery of highly-similar sequences while minimizing false haplotypes across a broad range of relative frequencies, read depths and platforms. This should allow users to maximize information extraction while minimizing biases in their downstream analyses and conclusions. In addition, the full SeekDeep suite of tools for pre- and post-processing will speed clustering optimization and provide high-quality and interpretable haplotype data for further analysis.

Code availability.

Source code for the current stable release of SeekDeep can be found at **https://github.com/bailey-lab/SeekDeep** and full usage and tutorials can be found at the SeekDeep website. For full install information see http://baileylab.umassmed.edu/SeekDeep/installingSeekDeep

Availability of Data and materials.

The *in vitro* data can be found via their original publications. The simulation raw data and the *P. falciparum* Illumina MiSeq data can be found at http://baileylab.umassmed.edu/data/SeekDeepPaperData.

MATERIALS AND METHODS

Overview of the SeekDeep Suite.

SeekDeep is a software suite written in C++ centered around *de novo* clustering providing rapid sample and input sequence preprocessing, and postprocessing sample and population summaries for further downstream analysis. SeekDeep can be utilized with most major sequencing technologies, including Ion Torrent, 454, and Illumina, to swiftly analyze numerous samples and amplicons (**Figure 2.22**). SeekDeep provides start-to-finish workflow from raw sequence files to population-level clustering and tabular and graphical summaries. SeekDeep is freely available under the GNU Lesser General Public License v3.0 and is actively developed on github (**https://github.com/bailey-lab/SeekDeep**) while usage and details on the program can be found at the SeekDeep website (**http://baileylab.umassmed.edu/SeekDeep/**). SeekDeep has three main components, extractor, qluster, and processClusters, that are central to generating clustering results, and an additional component, popClusteringViewer, to aid in viewing and sharing the results.

extractor: de-multiplexing and read filtering

The subprogram extractor is generalized to process 454 and Ion Torrent standard flowgram format (SFF) files and standard FASTQ files from any source. Extractor also demultiplexes samples and amplicons using a wide variety of barcode and primer schemes but can also operate on already demultiplexed data (e.g. data that has been demultiplexed by standard Illumina pipelines). Like most extraction programs, SeekDeep includes typical tools for initial filtering based on read length, presence of primers, quality score metrics,

and/or presence of ambiguous bases (i.e. Ns). Extractor first separates reads based on sample barcodes handling a wide range of barcoding schemes that are commonly employed. Next, multiple or a single pair of forward and reverse primers are detected, demultiplexed and removed. Filtering is then done on per base quality scores, and on expected read lengths which can be set per primer set. Also, optional contamination filtering can be performed by supplying the sequences of target regions whereby sequences that differ drastically from these are removed.

See **http://baileylab.umassmed.edu/SeekDeep/extractor_usage** for full details on the options offered by extractor.

qluster: rapid and accurate clustering based on quality

At the core of the SeekDeep package is the qluster algorithm, which iteratively collapses amplicon reads based on pairwise global alignments (**Figures 2.23-2.24**). It leverages sequencer-generated quality values to discern likely true differences from sequencing errors as well as k-mer frequencies to filter out likely low abundance PCR errors. Although SeekDeep can process multiple amplicons at once, they are processed independently and haplotypes are not built or phased across different amplicons. The clustering process is summarized below.

First, reads lacking differences are collapsed into identical sequence clusters, which are then indexed for k-mers (default size 9). These initial identical clusters are then sorted based on the associated number of reads. An iterative comparison process is then undertaken with successive rounds of clustering allowing for an increasing number of differences to trigger the merger of two clusters. Majority-rule consensus of the smallest clusters are pairwise aligned and compared sequentially to the consensus of the largest

clusters to determine if they should be merged into one cluster or remain as two separate clusters (**Figures 2.23-2.24**). Once the clusters have all been initially compared and collapsed, if meeting threshold, the threshold for collapse is stepwise-raised to allow for more divergence in a subsequent iteration. At the end of each iteration majority-rule consensus are generated to represent each of the clusters. If consensus have changed due to the addition of new sequences, the clusters are again compared at the same error thresholds before advancing to the next iteration. The algorithm allows for flexibility not only in the number of iterations, but also in the threshold number and type of differences to collapse. Differences are classified as one-base indels, two-base indels, greater than two-base indels, low-quality mismatches, high-quality mismatches, and low k-mer frequency mismatches. In this way, the clustering is similar to operational taxonomic unit (OTU) percent identity clustering, but instead of counting all differences equally we are able to weigh the type and the quality of the difference before determining whether to merge clusters - an important feature for sequencing technology-aware clustering.

For clustering iterations, there are default collapse threshold profiles for 454, Ion Torrent, and Illumina, or a custom file can be supplied. The custom input parameter file allows the expert user to balance sensitivity, specificity, and speed for specific applications. The default profiles were used for all analyses in this paper. For a 454/Ion Torrent dataset, our standard error profile limits initial collapsing to sequences differing by single-base indels, given that the predominant errors in these datasets are small indels caused by homopolymer misestimation. On an Illumina dataset, which is unlikely to have erroneous single-base indels but more likely to have base miscalls, the default profile does not collapse on indels but allows more low quality mismatches. This framework makes the qluster algorithm highly extensible and adaptable to changing error profiles in updated or novel

sequencing platforms. In terms of applications, the ability to collapse to an exact number of differences allows for biologic questions to be concretely addressed. For example, settings could allow 2-3 mismatches when sequencing viruses like HIV to collapse the viral clouds, or settings could be used to not allow a single high quality difference when searching for point mutations in the domain of a gene associated with drug resistance.

Differentiating mismatches with quality

The quality of any mismatch is determined by assessing the quality scores of the two mismatching bases in the pairwise alignment between clusters and the quality of the neighboring bases in the region (Altshuler et al. 2000). A primary quality and a neighboring quality is calculated. For a mismatch to be considered high-quality it must exceed the set thresholds for both of these quality values. The number of neighboring bases included can be changed; the default value is 2, which includes 2 bases upstream and downstream for a total of 4 neighboring bases examined. If a mismatch is determined to be a high quality error, its k-mer frequency is also checked to determine if the mismatch is in a low frequency k-mer. To calculate this, the mismatched base is centered in odd number length k-mer (defaulting to 9). Next, the previously indexed k-mers are checked to determine if mismatched centered k-mer has a low frequency – either as user defined or as a percentage of total reads. The k-mer cutoff defaults to 1 read, so if the k-mer occurs only once in the sample read set it is counted as a low frequency error. The k-mer position within the sequence can also be taken into account and helps to improve the filtering when repeats are present.

Homopolymer indel weighting for 454 and Ion Torrent

In the Ion Torrent and 454 technologies, the most common errors are indels in homopolymers. Thus, for homopolymers, indels are weighted to count less than other indels rather than separately categorizing them. Weighting incorporates the length by taking the size of the indel and dividing by the average size of both homopolymer runs. For example, a single-base indel found in a homopolymer of 4 bases (meaning one read has 4 bases and the other has 3 bases), the indel weight will be counted as 1/3.5 instead of 1.

Chimera Detection

After clustering, the resultant haplotypes can be examined for likely chimeras that may have resulted from PCR (**Figure 2.22b**). If replicates are available, then potential chimeras not appearing in all replicates will be removed. However, chimeras are often reproducible (Haas et al. 2011) which requires additional checks. This is accomplished by pairwise comparison of all the putative haplotypes from qluster checking to see if any cluster could be the result of a composite of two other clusters, which is similar to other approaches (Quince et al. 2011; Callahan et al. 2016). Since, by definition, parental haplotypes contributing to a cluster must preexist for a chimera to form from them, we normally require that the parents are of equal or greater abundance relative to the potential chimera. By default, chimeras are called when both parents are at least 2-fold greater in abundance (user definable). This is a conservative approach to minimize false discovery that prioritizes removal of artifactual chimeras at the cost of potentially excluding low abundance biologic recombinants, but for most applications chimeras tend to be more numerous. To minimize the loss of true biologic haplotypes in population analyses, we have implemented an option in our population clustering to check if a cluster marked as possibly chimeric appears in

another sample as one of the dominant haplotypes. If such a sample is found, the haplotype in question is, in that sample, unlikely to be chimeric since ideally a chimera would have two parents greater in abundance than itself. In this case, the putative chimera can be recovered in the original sample where it was at low-abundance. It is important to note that this step may not recover all true haplotypes as they might never appear at a high abundance in another sample. Also, a low-level chimera could be reinstated as a true haplotype. As there is no optimal solution for defining chimeras, we recommend every effort should be made during the PCR step to decrease likelihood of chimera formation. Also, chimera removal should be carefully considered and tuned, preferably with adequate controls, for the specific biology and experiment conditions. Again, these are options and during the sample and population clustering step it is possible to keep all putative chimeras for further analysis or to apply other chimera detection methods.

OTU Clustering

SeekDeep also offers classical OTU clustering, which is slightly modified to be calculated by taking into account only errors not characterized as low k-mer frequency or low quality mismatches and optionally weighing indels in homopolymers less when analyzing 454 and Ion Torrent data. In this way, the percent identity calculated takes into account only likely biological differences between sequences.

See **http://baileylab.umassmed.edu/SeekDeep/qluster_usage** for full qluster usage information.

processClusters: Replicate and Population Comparisons.

The qluster algorithm removes sequencing error and low level PCR error, but rare high-abundance errors due to polymerase errors in early rounds of PCR amplification are not easily discriminated. Therefore, when available, the SeekDeep pipeline uses PCR replicates (independent parallel amplifications of biologic sample aliquots) to identify and remove such errors – as early PCR errors should occur only in a single replicate, while biologic differences should occur faithfully in all samples. To compare replicates, the clustering results from each PCR are pooled and clustered again using the qluster algorithm. After this cross-replicate clustering a replicate number cutoff is applied, which defaults to the number of replicates used; for example, if three replicates were analyzed, the default would require all 3 replicates contain a given haplotype. Though PCR replicates are recommended they are not required for SeekDeep to run.

Additionally, a cutoff for the fraction of total reads within the cluster can also be given for comparison; if the average fraction of a new cluster is not above the cutoff, the new cluster is removed. This cutoff defaults to 0.005 (0.5%), a generally conservative cutoff to minimize false haplotypes for the vast majority of experimental conditions, but can be set to more appropriate levels. For chimera filtering, if the majority of a cluster is made up of reads marked as possibly chimeric, it is also marked as chimeric and is removed by default. Final relative abundances for haplotypes are re-calculated after cutoffs have been applied and when replicates are available the final abundances of a haplotype is calculated by averaging the abundances across the replicates.

In addition to replicate processing and applying final cutoffs, processClusters can also assess the haplotypes across samples to provide population-level statistics. Once each

sample has been processed, information is then collated across biologic samples within the defined population for each haplotype.

popClusteringViewer: viewing and manipulating final results

A web server has been added to the SeekDeep suite to aid in the visualization and exploration of final results; this can be very helpful with large sample sets. The viewer is interactive and allows rapid exploration of final consensus sequences and the population haplotypes. It can also be used to extract subsets of the data. The viewer can easily be run on an individual's computer and can also be broadcast over the internet to provide persistent access to additional individuals.

See **http://baileylab.umassmed.edu/SeekDeep/popClusteringViewer_usage** for full usage information.

Performance Studies

To validate performance of the SeekDeep pipeline, we used two types of data. The first was simulated 454 and Illumina datasets. The second was actual PCR-amplified and sequenced (by Ion Torrent, 454, and Illumina) control mixtures of DNA from strains of several different pathogens to create mock mixed infections, which were collected from several previous studies and work in our own lab. We also used available mock bacterial communities. See below for a detailed description of these datasets.

Simulated Datasets

The 454 and Illumina simulated datasets were created to test theoretical limits of detection for SeekDeep and other popular programs. The 454 datasets were simulated with

454sim (Lysholm, Andersson, and Persson 2011) and Illumina datasets were created with ART (Huang et al. 2012). While a specific Ion Torrent simulator could not be found, the 454 simulator should provide results representative of Ion Torrent pyrosequencing given their similarities. An in-house program was used to generate the PCR error by simulating the rounds of PCR where a PCR error that occurred in an earlier round would appear at higher abundance than latter round errors, a feature not available in other PCR simulators. The program takes a starting DNA template amount, PCR error rate, a fasta file with relative abundances for reference haplotypes to simulate, and the number of rounds to simulate. For these simulations we used 2,000 copies of starting DNA template, a PCR error rate of 3.5e-6 (representative of high-fidelity polymerases), and 30 rounds of PCR. Given the complexity of their formation, chimeras were not simulated.

Two mock haplotype mixtures were simulated to generate multiple test conditions:

- **Mock haplotype mixture 1 (Minor vs Major):** This mixture tests the ability of programs to discriminate minor haplotypes at various levels of divergence and abundance from a major abundant haplotype (**Figure 2.2a**); thereby assessing the likelihood of minor haplotypes being collapsed into the major as probable error. Specifically, we simulated seven different haplotypes with increasing base mismatches (decreasing % identity) of 1 (99.7%), 2 (99.4%), 3 (99.1%), 4 (98.8%), 6 (98.2%), 8 (97.6%), and 13 (96.1%) from the major haplotype, with no shared mismatches between minor haplotypes to create distances always greater to other minor haplotypes than to the major haplotype, e.g. the minor haplotype with 1 mismatch and the minor haplotype with 2 mismatches from the major haplotype are 3 mismatches away from each other. The relative abundance of the minor haplotypes were simulated at 10%, 5%, 2%, 1%, 0.5%, 0.25%, 0.1%, and 0.05% (**Figure 2.2b**).

- **Mock haplotype mixture 2 (Minor vs Minor Pairs with Varying Differences)**: This mixture examined the effect of divergence between minor haplotype pairs unrelated to the major haplotype (**Figure 2.2c**). For this, we simulated 15 different haplotypes, making one major abundant haplotype and 14 minor haplotypes. Each minor haplotype was paired with another closely related minor haplotype, and each haplotype in the pair differed by at least 15 mismatches from the other pairs or to the major haplotype. Pairs had a range of base mismatches (% identity) consisting of 1 (99.7%), 2 (99.4%), 3 (99.1%), 4 (98.8%), 6 (98.2%), 8 (97.6%), and 13 (96.1%) nucleotides. The relative abundances of the minor haplotypes were simulated at 5%, 2%, 1%, 0.5%, 0.25%, 0.1%, and 0.05% with the rest composed of the major haplotype. (**Figure S4d**).

For each mixture and minor haplotype abundance above, we generated simulated datasets with two replicate PCRs with 2,000-10,000 reads incrementing by 2,000 and at 50,000 reads (to test the extremes of coverage) for a total of 6 different read depths (equivalent throughout to nonredundant read or stitched-read coverage across the amplicon--or equivalently per base). Each of these conditions was simulated 10 times and the results were averaged to get the best estimate of program performance.

Known Control Mixture Datasets

Five different experimental *in vitro* control mixtures were analyzed spanning the common sequencing technologies; 454, Ion Torrent, and Illumina (**Table 2.2)**. This included data from a eukaryotic parasite (*Plasmodium falciparum*) and a mock microbiome. Specifically these were:

- ***Plasmodium falciparum* control mixtures, 454 and Ion Torrent***: Plasmodium falciparum* control mixtures from our labs were sequenced on Ion Torrent and 454

(**Table 2.2**). These pools contained three different amplicons: thrombospondin-related anonymous protein (*TRAP*) (**Figure 2.25**), apical membrane antigen 1 (*AMA1*) (**Figure 2.26**), and circumsporozoite protein (*CSP*) (**Figure 2.27**). The *AMA1* and *TRAP* samples had the same mixture of five strains: 40% K1, 30% 7G8, 15% Dd2, 10% RO33, and 5% V1/S and the *CSP* region had a mixture of 40% K1, 30% 7G8, 20% DD2, and 10% RO33 (**Figures 2.25-2.27**).

- ***Plasmodium falciparum* control mixtures, Illumina MiSeq**: Additionally, twenty-eight different regions, including vaccine candidates and drug resistance genes, were PCR amplified and sequenced with 2x250 paired-end Illumina MiSeq from a control mixture of *Plasmodium falciparum* (**Table 2.2)**. The mixture consisted of the following strains and relative abundances; 3D7 (~79%), HB3 (~7%), 7G8 (~7%), and DD2 (~7%). These targets included multiple probes in important vaccine candidate regions in *AMA1*, *CSP*, and merozoite surface protein 1 (*MSP1*). Also known drug resistance or associated loci were targeted including apicoplast ribosomal protein S10 (*ARPS10*), multidrug resistance protein 1 (*MDR1*), multidrug resistance protein 2 (*MDR2*), kelch13 (*K13*), protein phosphatase (*PPH*), Cytochrome b (*CYTB*), dihydrofolate reductase thymidylate synthase (*DHFR-TS*), and dihydropteroate synthase (*DHPS*) (**Figure 2.28)**.
- **Mock Microbiome:** Previous mock microbiome datasets by Salipante et al. 2014 were analyzed consisting of Illumina paired-end sequencing of the V1 region of the 16S coding region with 3 PCR replicates (Salipante et al. 2014). This mock microbiome mixture contains 20 species, but due to highly similar copies within each species the number of expected haplotypes at one-base resolution for the V1 region is 47. Twenty of these haplotypes are only one base pair different from another haplotype. The 3 PCR replicates were deeply sequenced with approximately 800,000 reads each. To analyze

data at more commonly assessed read depths (MacIntyre et al. 2015; Mideo et al. 2016; Lin et al. 2015) the replicates were downsampled to depths between 2,000-20,000 increasing by intervals of 2,000. Each read depth was sampled 10 times each for all 3 PCR replicates which generated a total of 300 different randomly sampled datasets.

- **Epstein-Barr Virus (EBV) and Human immunodeficiency virus type 1 (HIV) controls**: To provide a broader set of biologic examples, we also examined available viral controls of amplicon sequencing consisting of a previous mock HIV mixture (Seifert et al. 2016) and a mock EBV mixture from our lab. The HIV dataset had 5 strains mixed together; 89.6 (10%), HXB2 (14%), YU2 (16%) , NL4-3 (24%) and JR-CSF (36%). The mixture was sequenced 5 different times, two of the replicates were chosen and due to the great depth (>600,000) were each downsampled to 10,000 reads 10 times each for a total of 20 randomly sampled datasets. The EBV dataset was mixtures of an EBV type 1 strain and an EBV type 2 strain with frequencies ranging from 1% to 90% and a monoclonal sample of the type 1 strain. See **Table 2.2** for more details.

MED (version 2.1), DADA2 (version 1.0.3), UNOISE (USEARCH version 9.2), and SeekDeep (version 2.4.0) were each run on the datasets with default or recommended parameters. The program ShoRAH (Zagordi et al. 2011) (version 1.1.0) was used on the viral datasets to represent a standard program for viral analysis. DADA2 and UNOISE have their own chimera detection program; MED and ShoRAH do not have a chimera-detection utility, so our own chimera detection was applied to the final results produced by MED and ShoRAH to make the results comparable. Each program has a different output format from which the consensus sequences and relative abundances of final clusters were extracted. The expected abundance of pooled species for each dataset was determined by aligning

raw reads to reference sequences for that dataset. This calculation was performed because mock mixtures are manually produced in the lab, making the targeted mixture frequencies approximate. Common sources of experimental error arise by pipetting inaccuracy and imperfect amplification of the initial low abundance template leading to the introduction of random noise during the early rounds of PCR. Final clustering results were compared to the expected reference sequences and to determine which references were identified. For the paired-end Illumina data, sequences were stitched together with the program FLASH v1.2.11 (Magoc and Salzberg 2011).

To evaluate performance of each program we determined the number of expected haplotypes recovered - especially one-off haplotypes - and how well their abundances were predicted. We also determined the number and abundances of false haplotypes created. Recovery was calculated as the number of haplotypes exactly matching expected haplotypes divided by the total number of haplotypes expected. The haplotype recovery for MED, DADA2, and UNOISE was calculated based on each replicate separately, while SeekDeep's haplotype recovery was calculated if it found the expected haplotype in both replicates for a sample, as this is its default. Thus, SeekDeep's haplotype recovery is conservative relative to the other programs given that a haplotype must be present in both replicates to be counted as recovered.

All analyses and program comparisons were run on an Ubuntu 14.04 server with 64 2.4-GHz AMD processor cores and 512 gigabytes (GB) of RAM to allow parallelization of all simulations and *in vitro* datasets. For SeekDeep, all analyses presented could also be run individually on a laptop, a Macbook Pro with 16GB of RAM and a 4-core 2.4 GHz Intel i7 processor.

Tables

Table 2.1: Full Mock Microbiome Results

Program	Replicate	True Haplotypes Predicted	True Haplotypes Expected	Recall
UNOISE	1	44	47	93.62
UNOISE	2	44	47	93.62
UNOISE	3	43	47	91.49
MED	1	47	47	100
MED	2	47	47	100
MED	3	47	47	100
DADA2	1	46	47	97.87
DADA2	2	46	47	97.87
DADA2	3	46	47	97.87
SeekDeep	1	47	47	100
SeekDeep	2	47	47	100
SeekDeep	3	47	47	100

Table 2.2: *In vitro* control datasets

Dataset Amplicon	Technology	Read Depth*	Sample Number	Replicate**	Read Length	Region Length	Unique Haplotype Number	Range of Haplotype base differences (% identity) ***
PfTRAP	454	812 - 987	1	2x	345	345	5	1 (99.7%) - 7 (97.9%)
PfAMA1	Ion Torrent	1,323 - 1,712	2	2x	494	494	5	2 (99.1%) - 12 (94.9%)
PfCSP	Ion Torrent	1,054 - 6,403	4	2x	319	319	4	2 (99.3%) - 9 (97.2%)
Various *P. falciparum* targets****	Illumina MiSeq	614 - 4,497	28	none	2x250	330-403	2-4	1 (99.7%) - 17 (95.3%)
Microbiome 16S-V1	Illumina MiSeq	584,575 - 899,804	1	3x	2x250	280	47	1 (99.6%) – 101 (63.9%)
EBV	Illumina MiSeq	342 - 1,350	6	2x	2x250	372	2	20 (92.6%)
HIV	Illumina MiSeq	10,000	20	2x	2x250	206	5	2 (99%) - 5 (97.5%)

* Read depth equals number of stitched read pairs with minimum and maximum observed depths in the case of multiple samples and replicates.

** 2x = two independent PCRs; 3x = three independent PCRs, or none = no replicate (single PCR) done

*** Number of differences are enumerated and followed by the corresponding percent identity, the range is shown when there are more than 2 unique haplotypes

**** Summary of the 28 targets here, see **Table 2.3** for details for each target

Table 2.3: *In vitro P. falciparum* Illumina control datasets

Dataset Amplicon	Technology	Read Depth*	Sample Number	Replicate**	Read Length	Region Length	Unique Haplotype Number	Range of Haplotype base differences (% identity) ***
PfAMA1_0	Illumina MiSeq	1,362	1	none	2x250	389	4	1 (99.7%) - 4 (99%)
PfAMA1_1	Illumina MiSeq	1,016	1	none	2x250	389	4	9 (97.7%) - 14 (96.4%)
PfAMA1_2	Illumina MiSeq	3,220	1	none	2x250	395	4	9 (97.7%) - 14 (96.5%)
PfAMA1_3	Illumina MiSeq	2,171	1	none	2x250	392	4	2 (99.5%) - 9 (97.7%)
PfAMA1_4	Illumina MiSeq	1,221	1	none	2x250	359	4	2 (99.4%) - 6 (98.3%)
PfAMA1_5	Illumina MiSeq	936	1	none	2x250	403	4	1 (99.8%) - 8 (98%)
PfAMA1_6	Illumina MiSeq	1,642	1	none	2x250	387	4	1 (99.7%) - 7 (98.2%)
PfARPS10_2	Illumina MiSeq	2,159	1	none	2x250	330	2	1 (99.7%)
PfCSP_1	Illumina MiSeq	926	1	none	2x250	354	4	2 (99.4%) - 9 (97.5%)
PfCYTB_2	Illumina MiSeq	4,303	1	none	2x250	361	2	1 (99.7%)
PfDHFR-TS_0	Illumina MiSeq	4,497	1	none	2x250	371	4	1 (99.7%) - 3 (99.2%)
PfDHFR-TS_2	Illumina MiSeq	3,165	1	none	2x250	360	4	1 (99.7%) - 3 (99.2%)
PfDHFR-TS_3	Illumina MiSeq	1,365	1	none	2x250	349	2	1 (99.7%)
PfDHPS_5	Illumina MiSeq	2,946	1	none	2x250	346	3	1 (99.7%) - 2 (99.4%)
PfDHPS_6	Illumina MiSeq	1,433	1	none	2x250	388	2	1 (99.7%)
PfK13_8	Illumina MiSeq	614	1	none	2x250	351	2	2 (99.4%) - 3 (99.1%)
PfMDR1_0	Illumina MiSeq	2,347	1	none	2x250	358	3	2 (99.4%)
PfMDR1_1	Illumina MiSeq	2,059	1	none	2x250	391	4	1 (99.7%) - 3 (99.2%)
PfMDR1_11	Illumina MiSeq	3,075	1	none	2x250	360	3	1 (99.7%) - 2 (99.4%)
PfMDR1_12	Illumina MiSeq	3,846	1	none	2x250	373	3	1 (99.7%) - 2 (99.5%)
PfMDR1_13	Illumina MiSeq	2,067	1	none	2x250	347	2	1 (99.7%)
PfMDR1_2	Illumina MiSeq	1,960	1	none	2x250	402	2	1 (99.8%)
PfMDR2_3	Illumina MiSeq	3,159	1	none	2x250	359	2	1 (99.7%)
PfMDR2_5	Illumina MiSeq	2,292	1	none	2x250	354	4	2 (99.4%) - 2 (99.4%)
PfMSP1_2	Illumina	1,051	1	none	2x250	364	4	1 (99.7%) - 17

	MiSeq							(95.3%)
PfPPH_1	Illumina MiSeq	2,770	1	none	2x250	403	2	1 (99.8%)
PfPPH_10	Illumina MiSeq	2,384	1	none	2x250	358	2	1 (99.7%)
PfPPH_11	Illumina MiSeq	1,171	1	none	2x250	384	4	1 (99.7%) - 3 (99.2%)

Figures

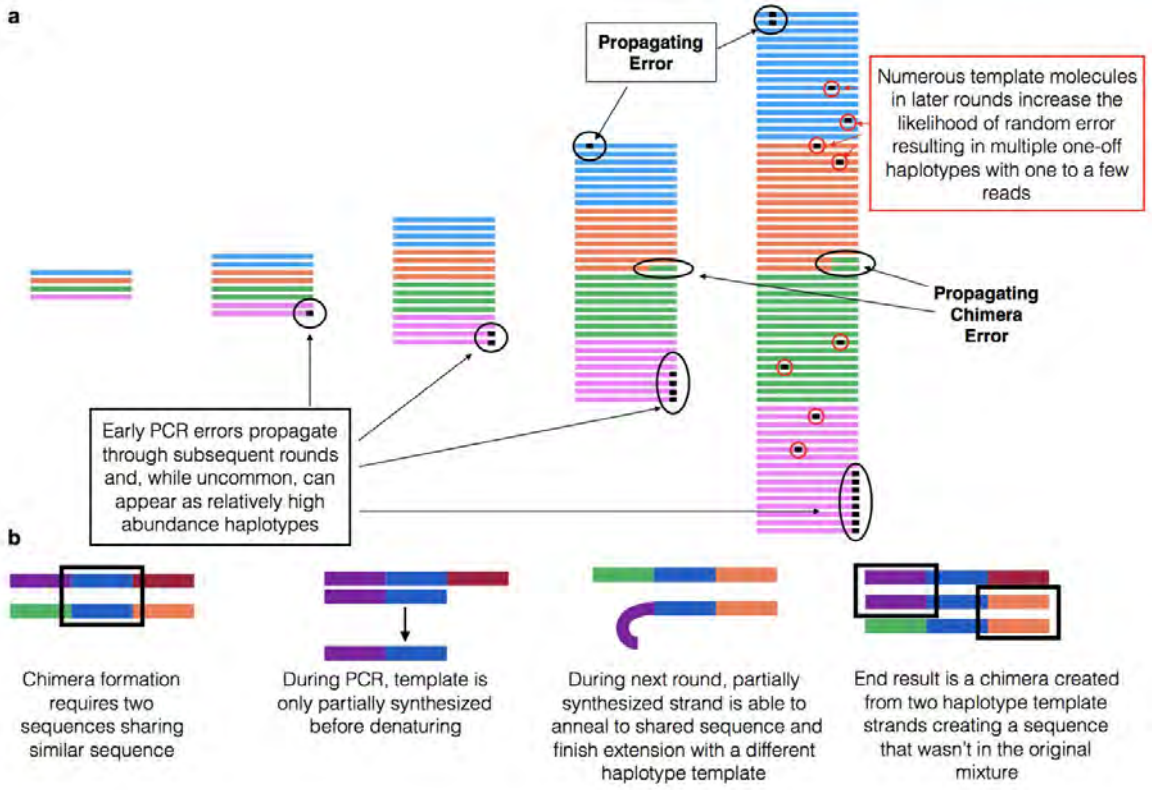


Figure 2.1: PCR and Sequencing Errors

Clustering of amplicon sequencing must contend with errors that occur during PCR and sequencing. **a)** Early round PCR errors can be difficult to identify because these propagate in subsequent rounds and can reach a relatively-high abundance. While usually uncommon, the degree of such high-abundance errors is highly dependent on the number of initial target DNA copies in the PCR. Thus, experiments that utilize nested PCR to amplify low DNA concentration samples are particularly prone to high-abundance errors. Low-abundance errors that occur in later rounds of amplification may be numerous but are more easily identified and removed. **b)** Another common problem in PCR, particularly when co-amplifying highly-similar sequences, is the creation of chimeras, which are formed when a partial PCR product re-anneals to a similar template creating a hybrid product.

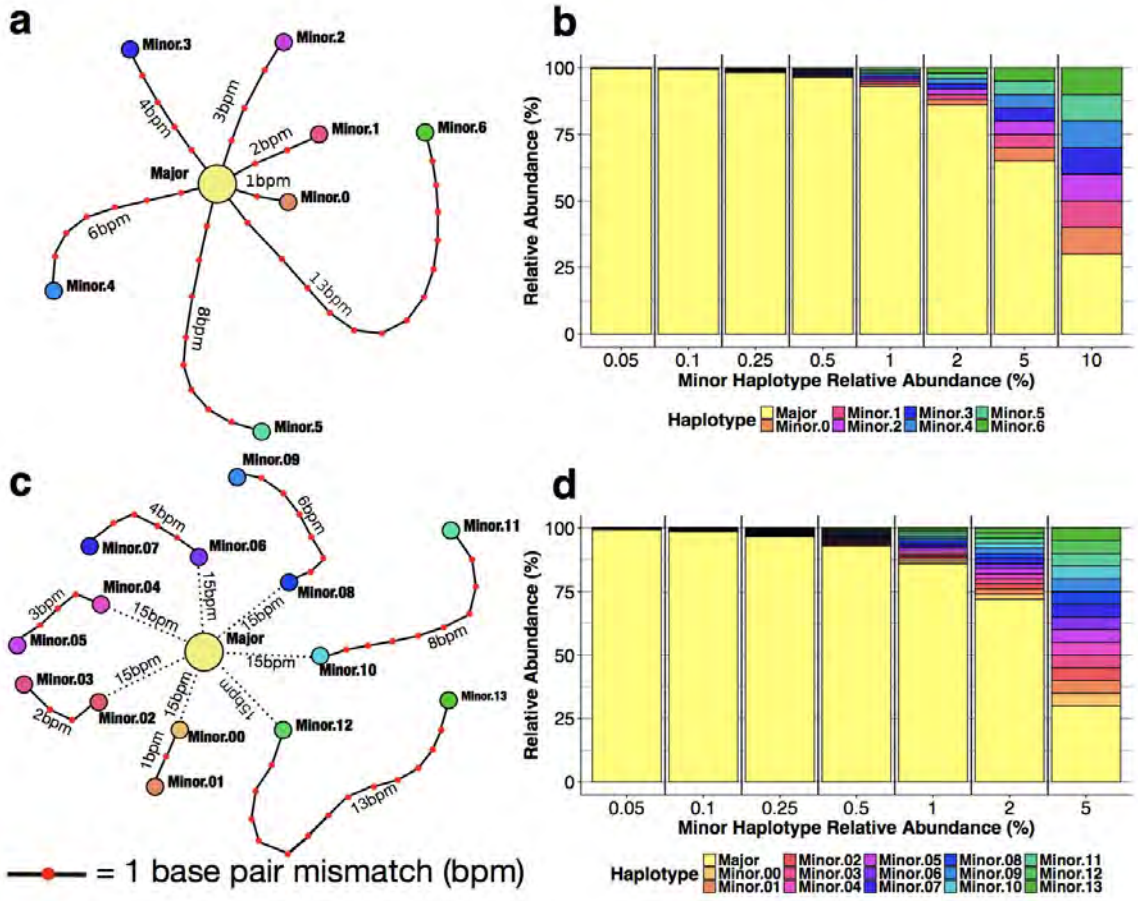


Figure 2.2: Simulated Mixtures

Two types of haplotype mixtures were simulated to assess performance. **a)** The first mixture tests discrimination of related low-abundant minor haplotype and highly-abundant major haplotype and is comprised of 7 minor haplotypes differing from the major haplotype by 1 to 13 differences. **c)** The second simulated mixture tests the ability to discriminate highly similar low-abundance haplotypes from each other. There are seven minor haplotype pairs differing by 1, 2, 3, 4, 6, 8, or 13 nucleotides. Between pairs and between the major haplotypes there are at least 15 nucleotides (all red dots not shown). **b)** The 8 different abundances at which haplotypes in panel **a)** were simulated. **d)** The 7 different abundances at which haplotypes in panel **c)** were simulated. Each was simulated 10 times at a variety of read depths. The number of red nodes between haplotypes is the number of base pair mismatches (bpm) differentiating them.

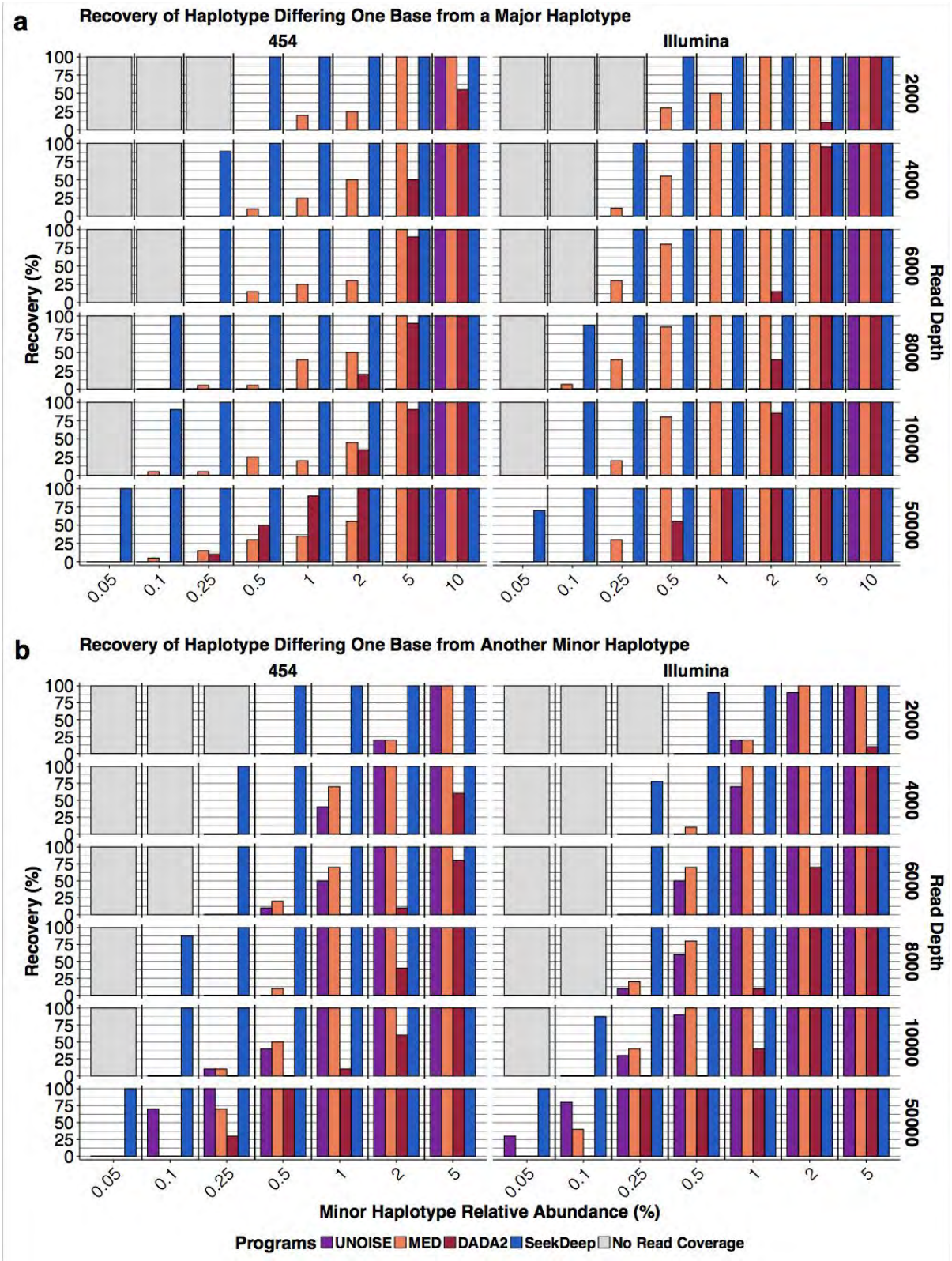

a
Recovery of Haplotype Differing One Base from a Major Haplotype
454
Illumina
Recovery (%)
Read Depth
2000
4000
6000
8000
10000
50000
0.05
0.1
0.25
0.5
1
2
5
10
b
Recovery of Haplotype Differing One Base from Another Minor Haplotype
454
Illumina
Recovery (%)
Read Depth
Minor Haplotype Relative Abundance (%)
Programs UNOISE MED DADA2 SeekDeep No Read Coverage

Figure 2.3: Haplotype Recovery of Simulated Minor Haplotypes Differing by a Single Base

a) Recovery of the haplotype differing by a single-base from a major haplotype in the mixture described by **Figure 2.2a-b**. **b)** Recovery of the two minor haplotypes that are one-off from each other described in the mixture described by **Figure 2.2c-d**. For both panels**,** the y-axis represents the percent of simulations in which the haplotype differing by a single-base was detected and the x-axis represents the simulated expected abundance of the minor haplotype. Data is broken down by read depth (rows) and sequencing technology (columns), and bars are colored by program. Grey boxes at low-abundances represent combinations where the depth is not sufficient for reads to be observed for the minor haplotypes. For each minor haplotype abundance, there are 20 simulations from which DADA2, MED and UNOISE haplotype recovery was calculated as a percent of simulations in which the minor haplotype was detected. To best emulate real world situations in which a user would use SeekDeep to analyze replicates, we used paired simulations with the requirement that SeekDeep detect haplotypes in both simulations.

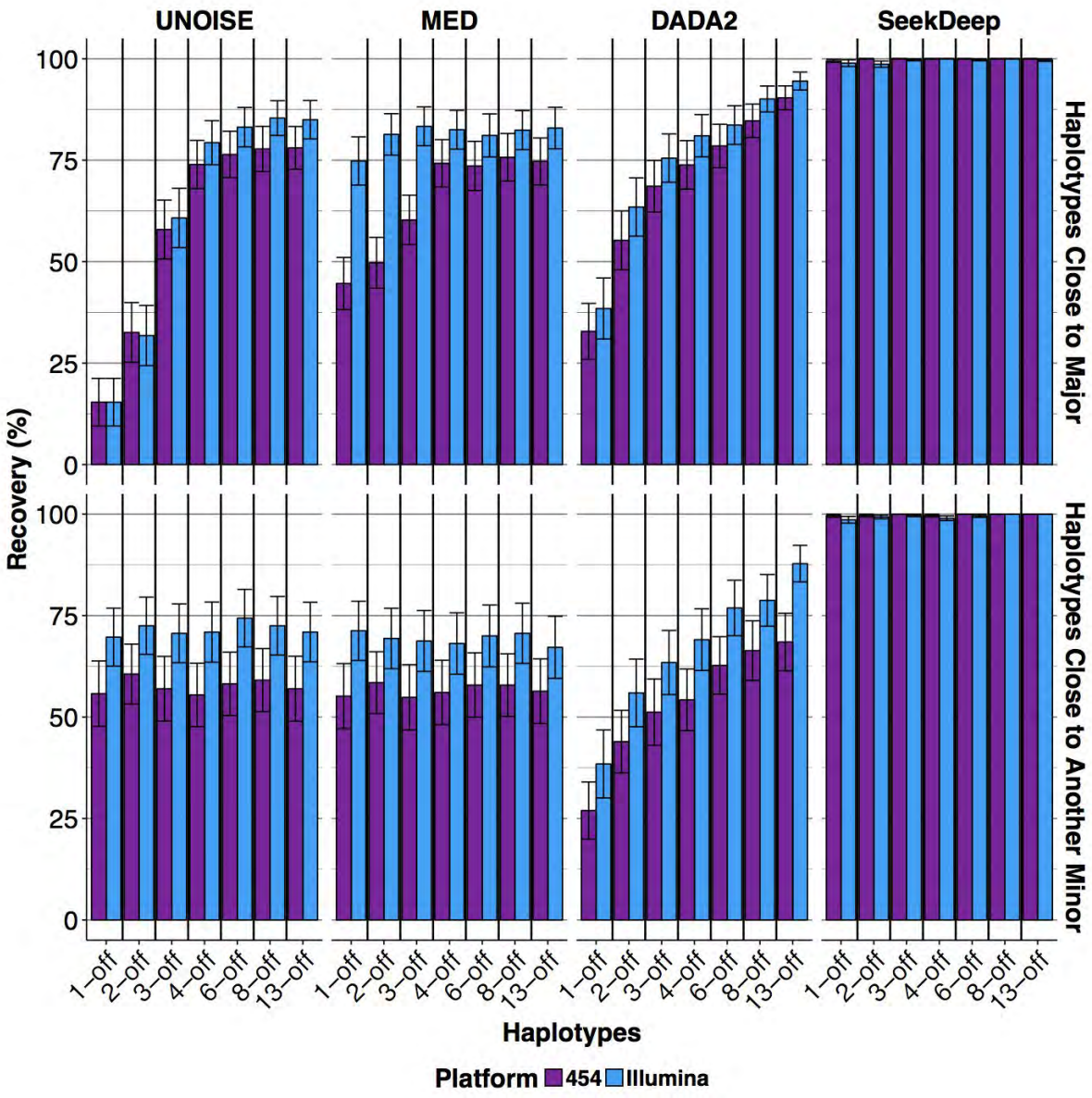


Figure 2.4: Haplotype Recovery of Simulation Data - Platform

The average haplotype recovery of the simulation datasets binned on technology and minor haplotype divergence for each program. The top row shows the average haplotype recovery of the minor haplotypes closely related to the major haplotype (**Figure 2.2a**), and the bottom is the average haplotype recovery of minor haplotypes close to another minor haplotype (**Figure 2.2c**). Error bars represent one standard error.

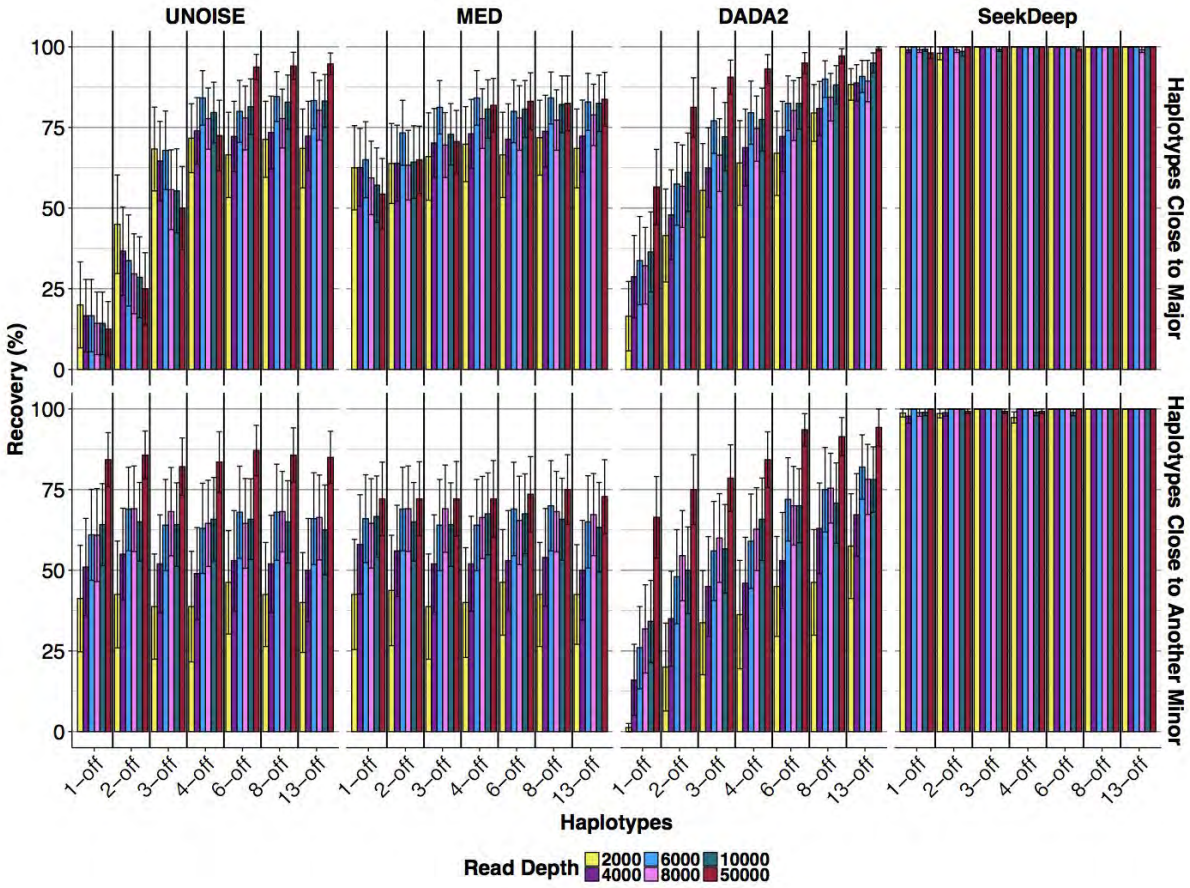


Figure 2.5: Haplotype Recovery of Simulation Data - Read Depth

The average haplotype recovery of the simulation datasets binned on simulated read depth and haplotype divergence for each program. The top row shows the average haplotype recovery of the minor haplotypes closely related to the major haplotype (**Figure 2.2a**) and the bottom is the average haplotype recovery of minor haplotypes close to another minor haplotype (**Figure 2.2c**). Error bars represent standard error.

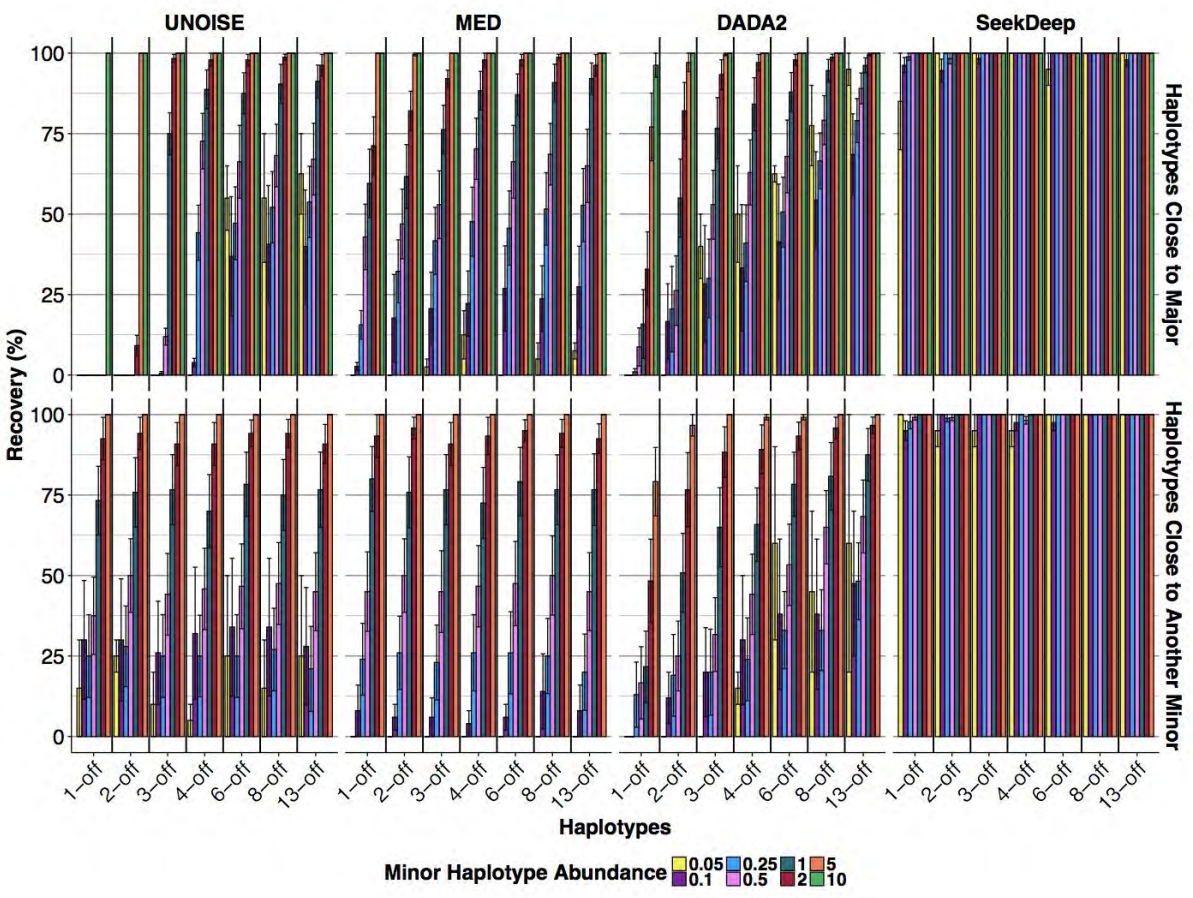


Figure 2.6: Haplotype Recovery of Simulation Data - Minor Haplotype Abundance

The average haplotype recovery of the simulation datasets binned on minor haplotype abundance and divergence for each program. The top row shows the average haplotype recovery of the minor haplotypes close to a major haplotype (**Figure 2.2a**), and the bottom is the average haplotype recovery of minor haplotypes close to another minor haplotype (**Figure 2.2c**). Error bars represent standard error.

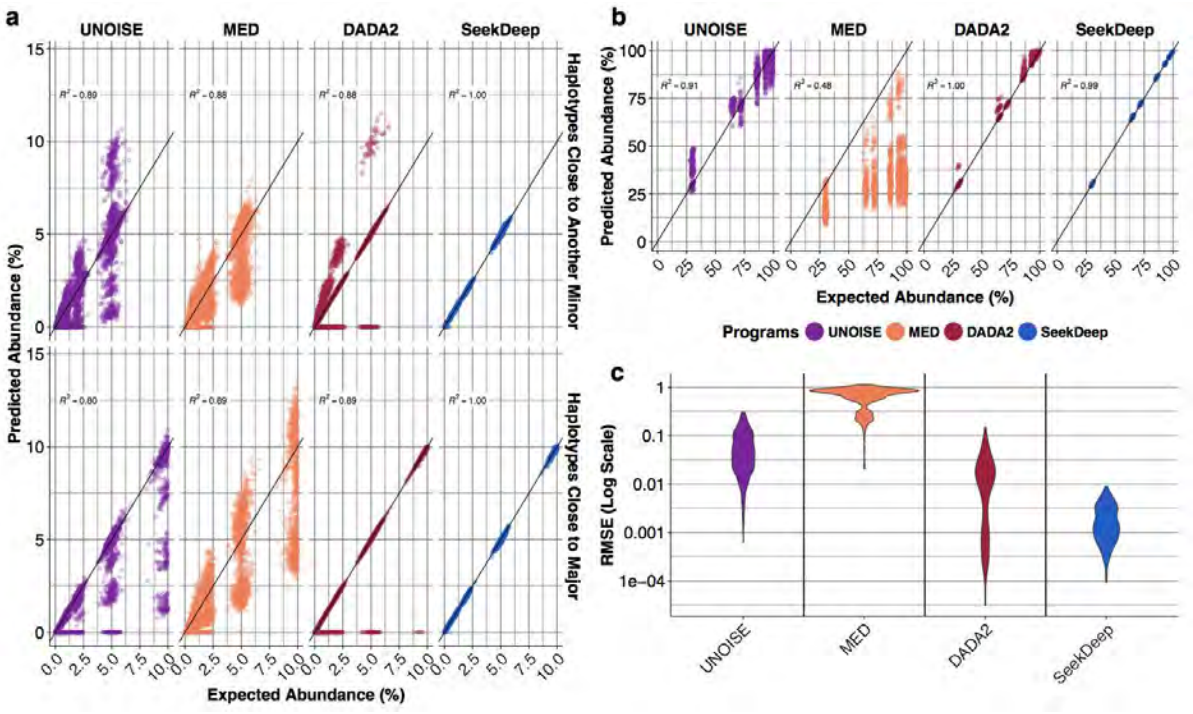
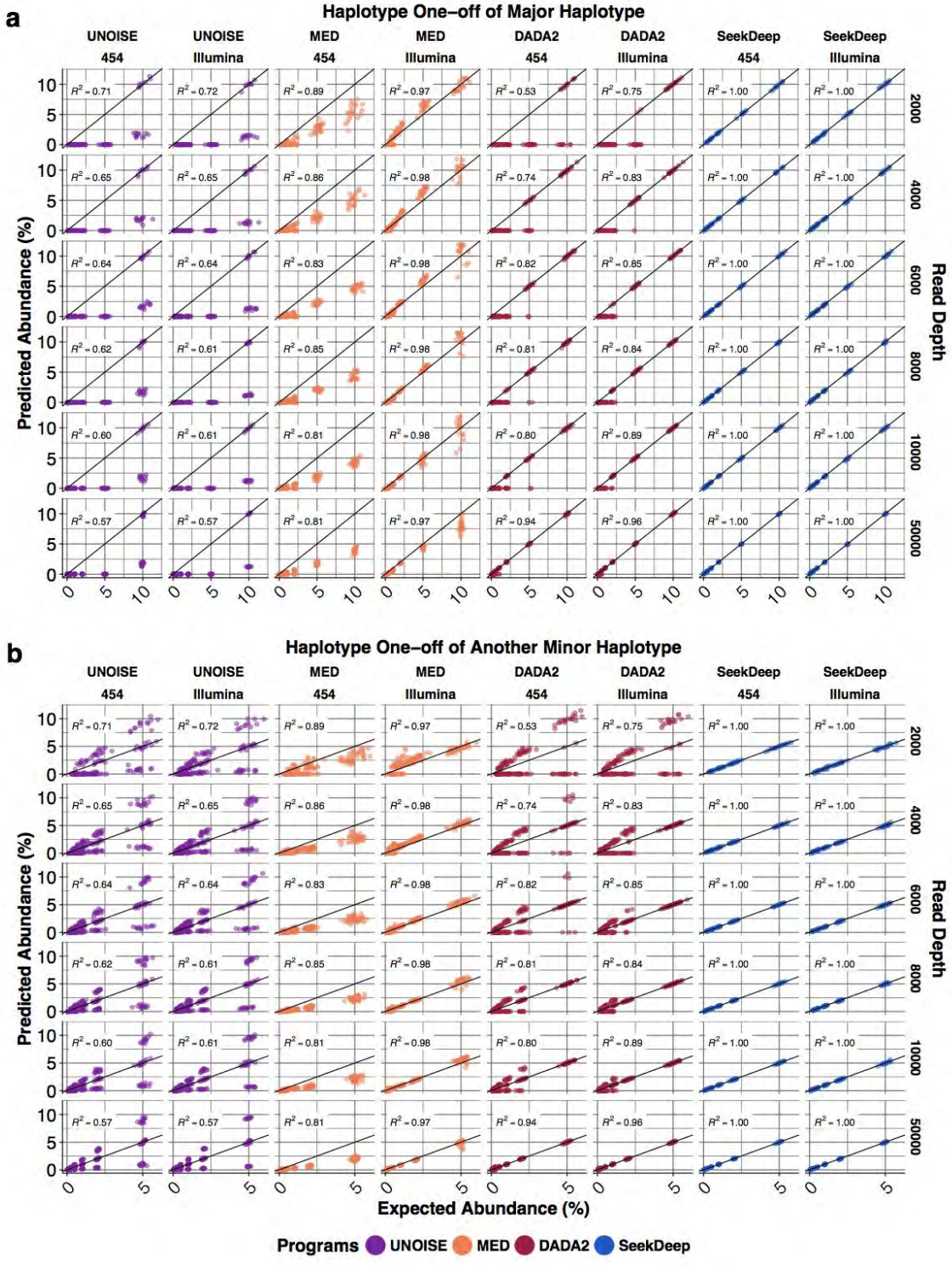
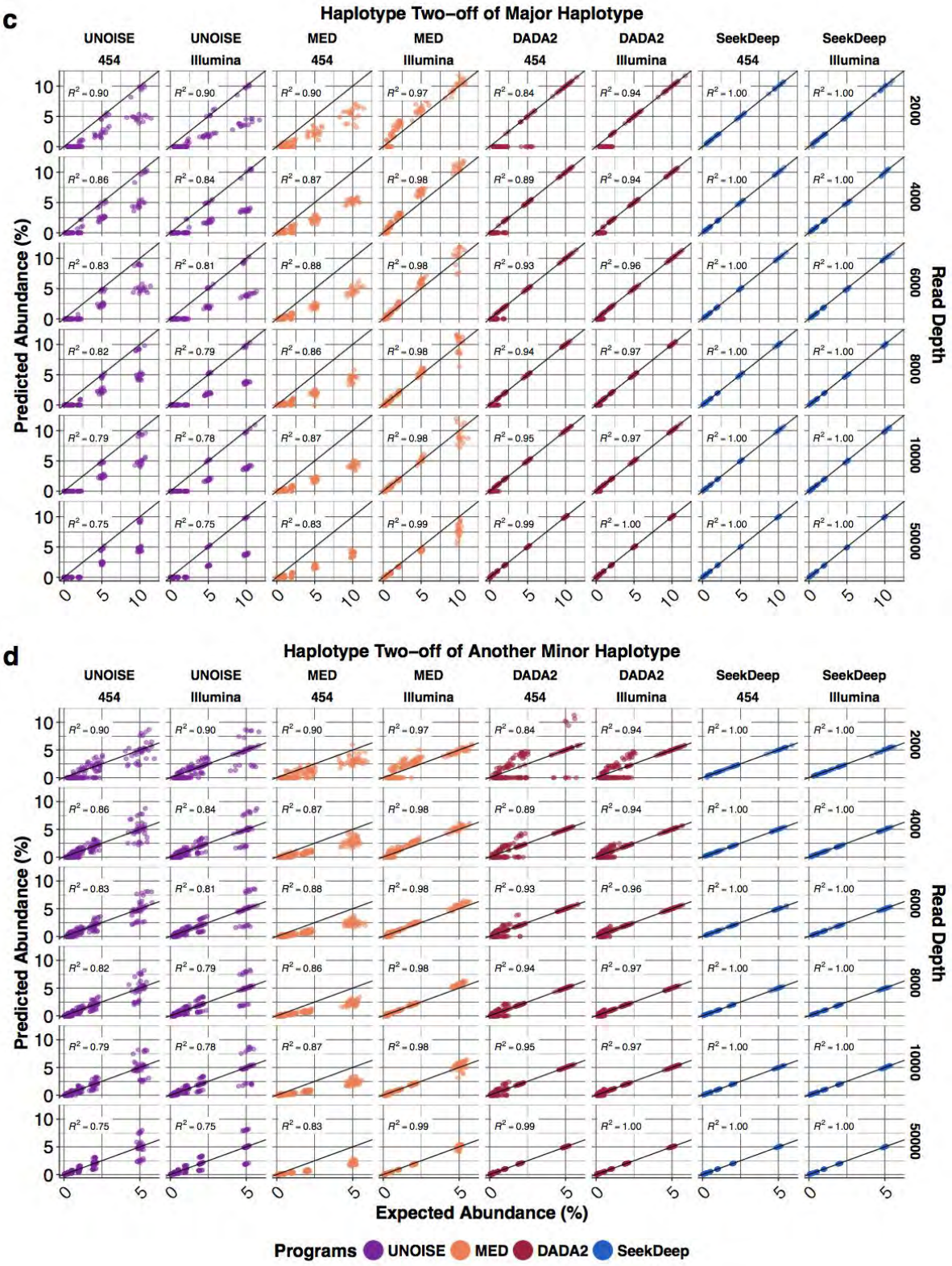


Figure 2.7: Predicted vs Expected Haplotype Abundances for Simulations

Panel **a)** is the plot of every simulated minor haplotype comparing each program's predicted abundance to the expected abundance based on direct read counts (**Figure 2.2** mixtures). Panel **b)** is the complementary plot of the major haplotypes for all simulations. **c)** A violin plot of the root mean squared error (RMSE) on the y-axis on a log scale for each program for all simulated datasets. For panels **a)** and **b),** the black line of identity for expected and predicted is shown. If points are above the line of identity the program is overestimating the abundance of the haplotype and if points are below the line the program is underestimating the abundance of the haplotype. The Spearman's correlation (R_2) is in the upper left corner of each plot.


a
Haplotype One−off of Major Haplotype
UNOISE 454
UNOISE Illumina
MED 454
MED Illumina
DADA2 454
DADA2 Illumina
SeekDeep 454
SeekDeep Illumina
Predicted Abundance (%)
Read Depth
2000
4000
6000
8000
10000
50000
b
Haplotype One−off of Another Minor Haplotype
Expected Abundance (%)
Programs UNOISE MED DADA2 SeekDeep


c
Haplotype Two−off of Major Haplotype
UNOISE 454
UNOISE Illumina
MED 454
MED Illumina
DADA2 454
DADA2 Illumina
SeekDeep 454
SeekDeep Illumina
Predicted Abundance (%)
Read Depth
2000
4000
6000
8000
10000
50000
d
Haplotype Two−off of Another Minor Haplotype
Expected Abundance (%)
Programs UNOISE MED DADA2 SeekDeep

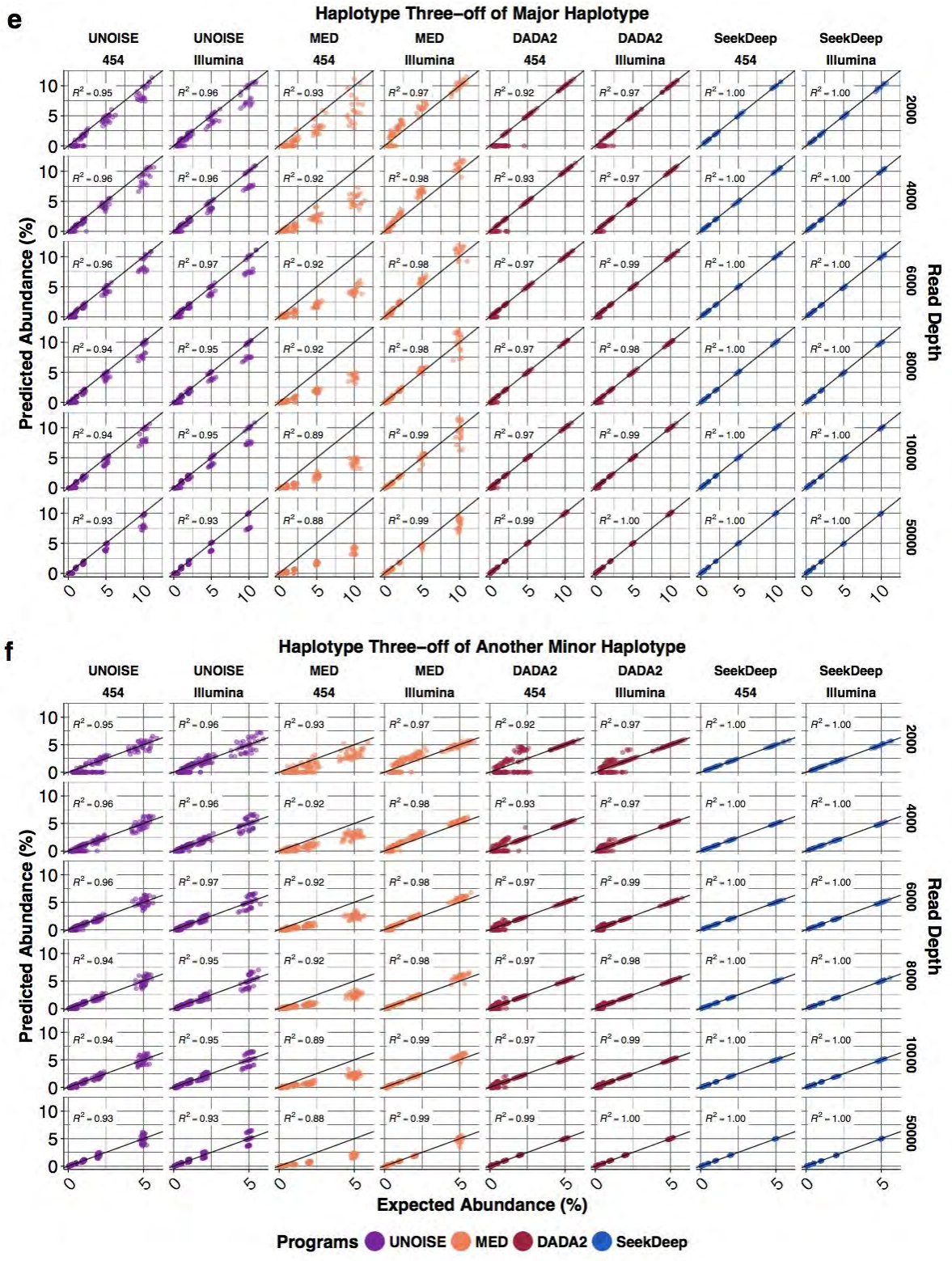

e
Haplotype Three−off of Major Haplotype
f
Haplotype Three−off of Another Minor Haplotype
UNOISE 454
UNOISE Illumina
MED 454
MED Illumina
DADA2 454
DADA2 Illumina
SeekDeep 454
SeekDeep Illumina
Predicted Abundance (%)
Expected Abundance (%)
Read Depth
2000
4000
6000
8000
10000
50000
Programs UNOISE MED DADA2 SeekDeep

Figure 2.8: Predicted vs Expected Haplotype Abundances for Simulations of Closely Related Haplotypes

The predicted vs expected abundances for known haplotypes differing by only one (**Panels a-b**), two (**Panels c-d**), or three (**Panels e-f**) bases is plotted to illustrate the effects of different read depths and technology for each program. Data points are colored by program. A diagonal black line is drawn to indicate perfect predicted for the expected abundance. Points above this line are overestimating haplotype abundance and points below this line are underestimating haplotype abundance. The Spearman's correlation has been placed in the upper left corner of each plot. Panels are **a)** haplotypes one mismatch off a major haplotype, **b)** haplotypes one mismatch off of another minor haplotype, **c)** haplotypes two mismatches off a major haplotype, **d)** haplotypes two mismatches off of another minor haplotype, **e)** haplotypes three mismatches off a major haplotype, and **f)** haplotypes three mismatches off of another minor haplotype.

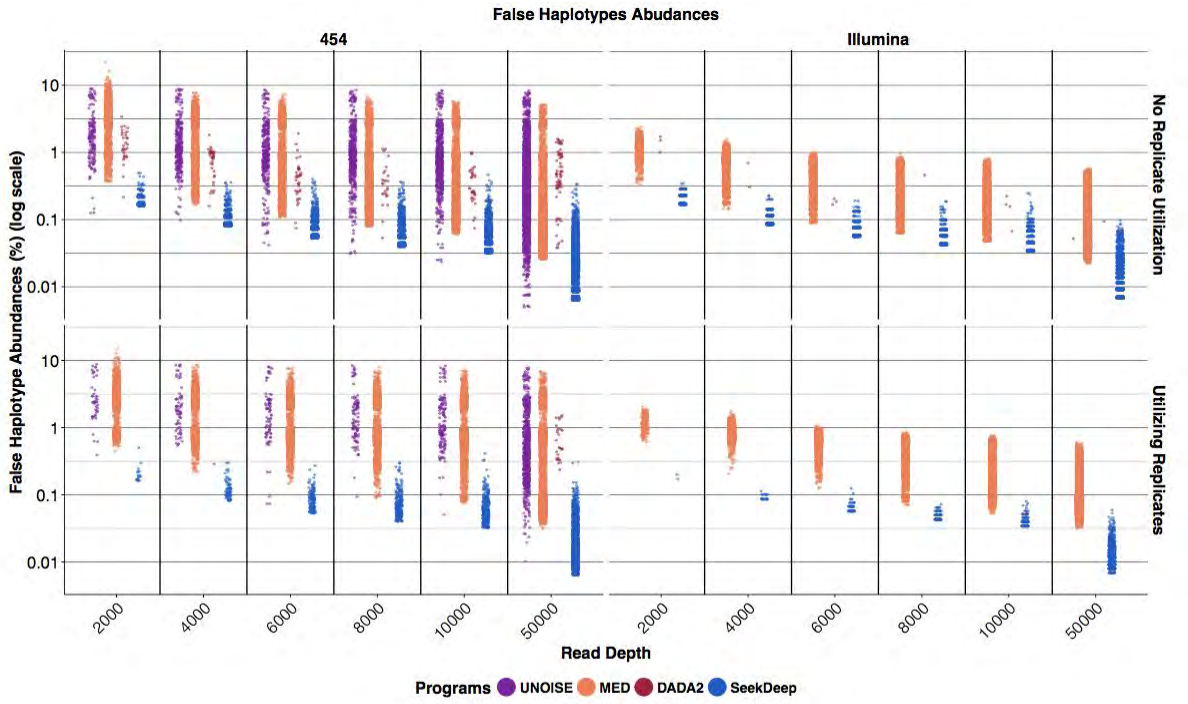


Figure 2.9: False Haplotype Abundances from Simulations

The relative abundances of predicted false haplotypes are binned by read depth (x-axis), technology (columns), and the use of replicates (rows). The y-axis is log scaled and is the relative abundance at which the false haplotypes were predicted.

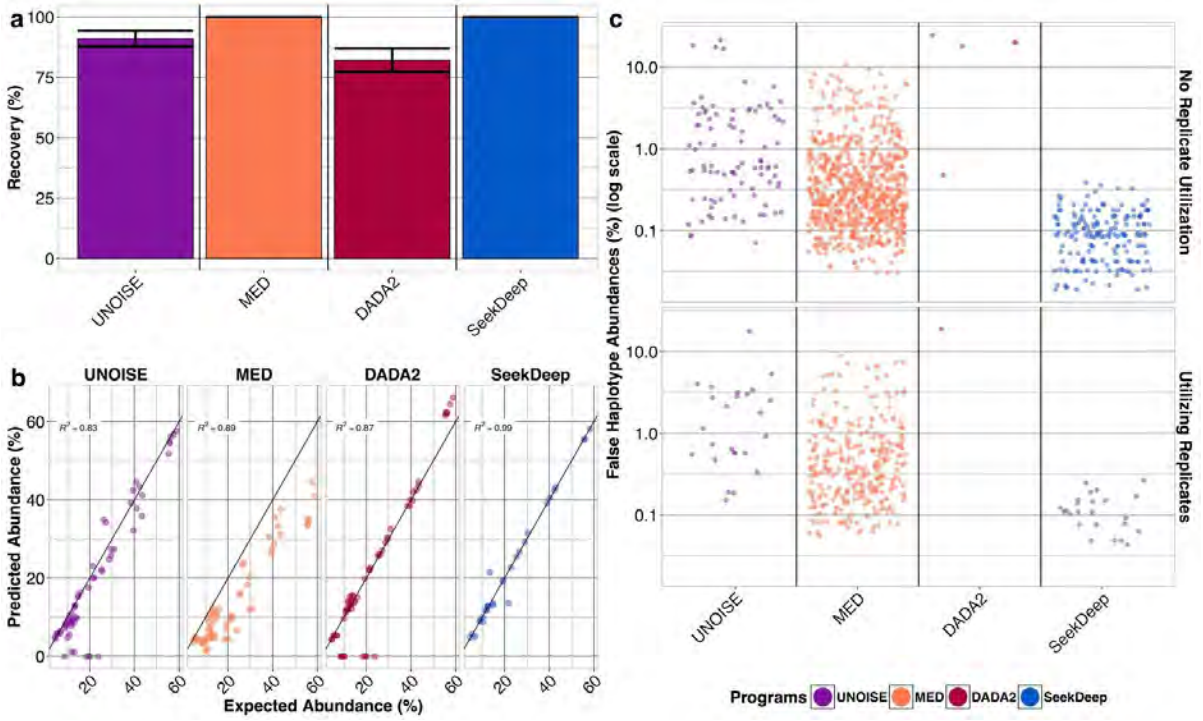


Figure 2.10: *In Vitro* Ion Torrent and 454 Mixtures Performance

a) The mean haplotype recovery for *in vitro* pyrosequencing samples with bars showing standard error. **b)** Predicted abundance (y-axis) estimated by the various programs is plotted against the expected abundance (x-axis). Deviation from the line of identity represents the error and is summarized by the correlation coefficient. **c)** False haplotypes are shown on a jitterplot to demonstrate their relative abundances and numbers. Results are shown per program and also by the effect of utilizing or not utilizing replicates (haplotypes are only accepted if they appear in both replicates).

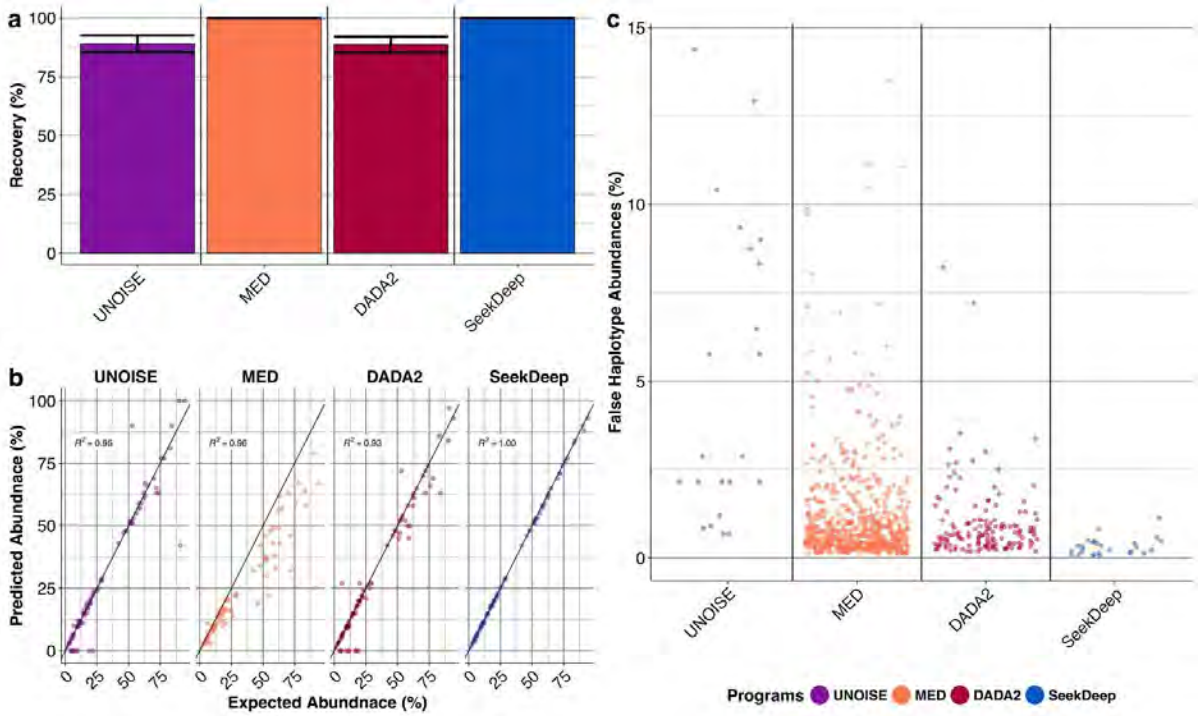


Figure 2.11: *In Vitro* Illumina *P. falciparum* Performance

a) The mean haplotype recovery for *P. falciparum in vitro* Illumina datasets with bars showing standard error. **b)** Predicted abundance (y-axis) estimated by the various programs plotted against the expected abundance (x-axis). Deviation from the line of identity represents the error and is summarized by the correlation coefficient. **c)** False haplotypes are shown on a jitterplot to demonstrate their relative abundances and numbers. No replicates were available for this dataset.

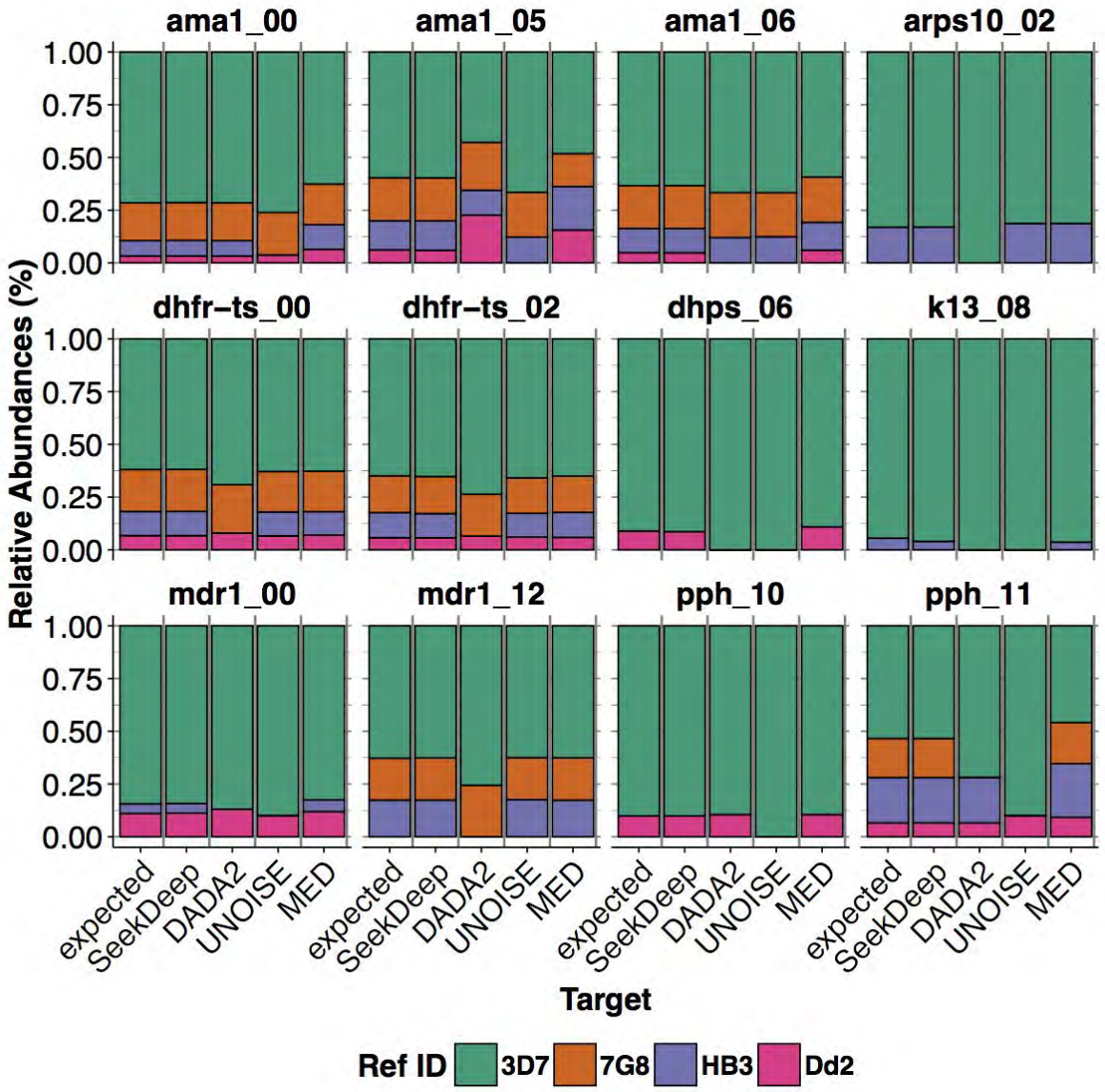


Figure 2.12: *In vitro* P. falciparum Illumina Mixtures Performance

The expected vs predicted abundances for all the target regions for when any of the programs failed to recover one of the expected reference haplotypes. The leftmost bar is the expected abundance based on direct mapping to the known reference and subsequent bars represent predicted abundances by program.

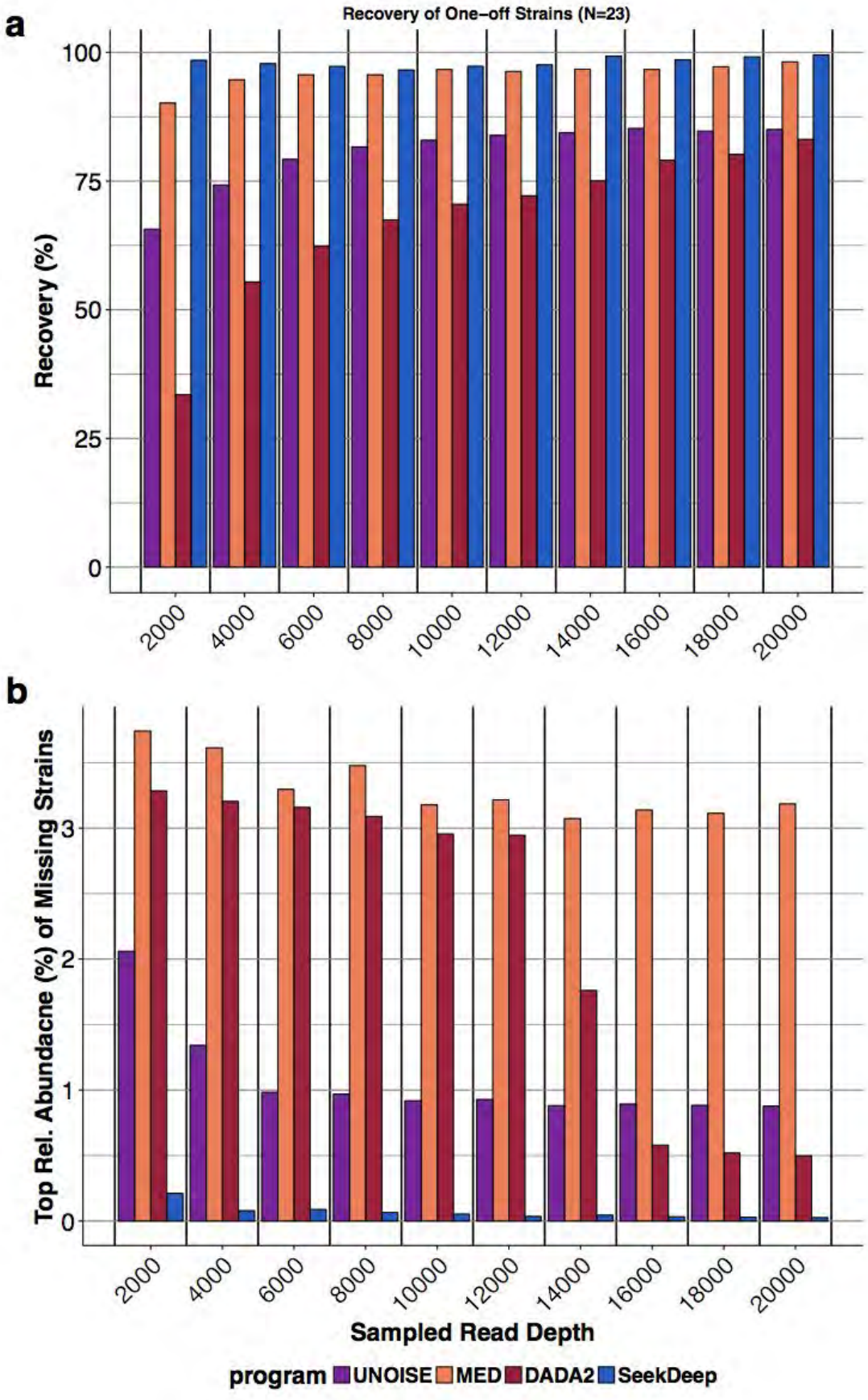

a
Recovery of One−off Strains (N=23)
Recovery (%)
100
75
50
25
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
b
Top Rel. Abundacne (%) of Missing Strains
3
2
1
0
Sampled Read Depth
program UNOISE MED DADA2 SeekDeep

Figure 2.13: Down-sampled Mock Microbiome Haplotype Recovery of Haplotypes Differing by One Base

a) Shows the haplotype recovery of the 23 haplotypes that are one-off from another haplotype in an overall mixture of 47 bacterial haplotypes which were down-sampled from the Salipante *et al*. 2014 data. Each of the three original replicates was down sampled randomly 10 times for each of 10 different read depths, which means each read depth has 30 randomly down sampled samples. **b)** A bar graph of the greatest observed abundance of missed one-off haplotype is shown for each program at each read depth.

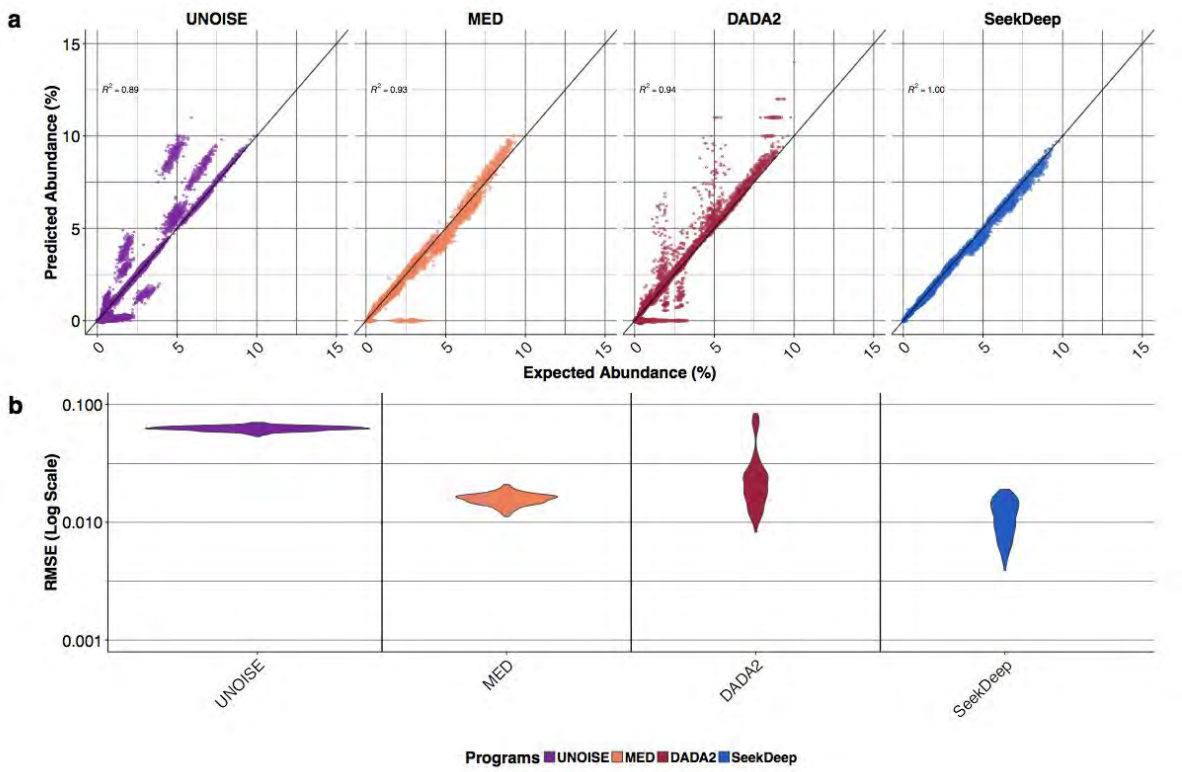


Figure 2.14: Down-sampled Mock Microbiome Predicted vs Expected Haplotype Abundances

a) A plot of predicted vs expected haplotype abundances is shown for each program for all down sampled datasets from Salipante *et al*. 2014. If the program predicted an abundance that equalled the expected abundance, points would fall on the depicted black line of identity. If a program overestimated the haplotype abundance, points would fall above the line. If a program underestimated the haplotype abundance, points would fall below the line. **b)** The log-scaled RMSE is shown as a violin plot for all down sampled data.

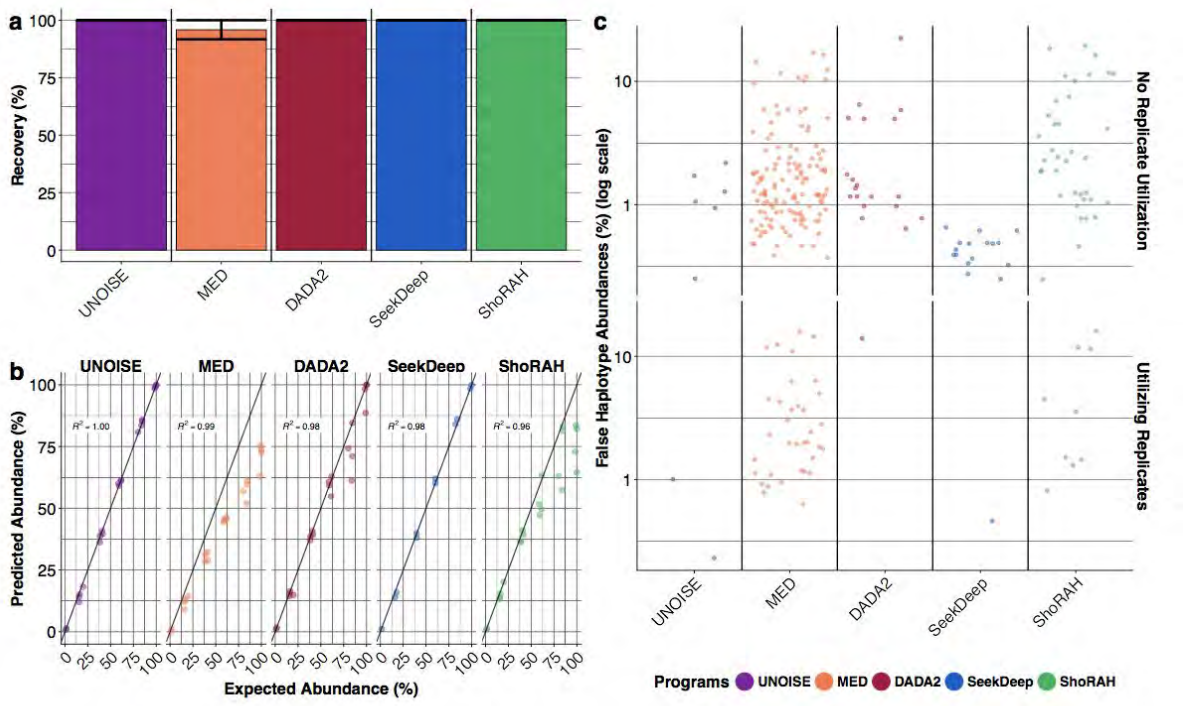


Figure 2.15: In Vitro EBV Illumina Performance

a) The mean haplotype recovery for the EBV datasets. **b)** Predicted abundance (y-axis) estimated by the various programs is plotted against the expected abundance (x-axis). Deviation from the line of identity represents the error and is summarized by the correlation coefficient. **c)** False haplotypes are shown on a jitterplot to demonstrate their relative abundances and numbers. Results are shown per program and also by the effect of utilizing or not utilizing replicates (haplotypes are only accepted if they appear in both replicates).

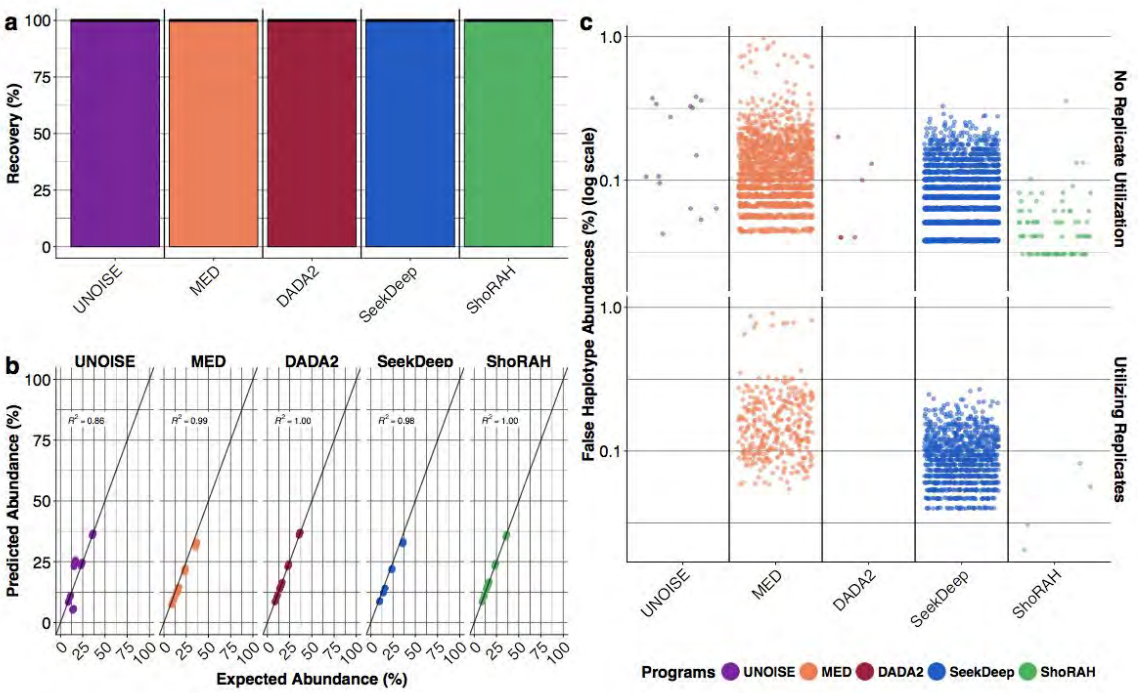


Figure 2.16: In Vitro HIV Illumina Performance

a) The mean haplotype recovery for the HIV datasets. **b)** Predicted abundance (y-axis) estimated by the various programs is plotted against the expected abundance (x-axis). Deviation from the line of identity represents the error and is summarized by the correlation coefficient. **c)** False haplotypes are shown on a jitterplot to demonstrate their relative abundances and numbers. Results are shown per program and also by the effect of utilizing or not utilizing replicates (haplotypes are only accepted if they appear in both replicates).

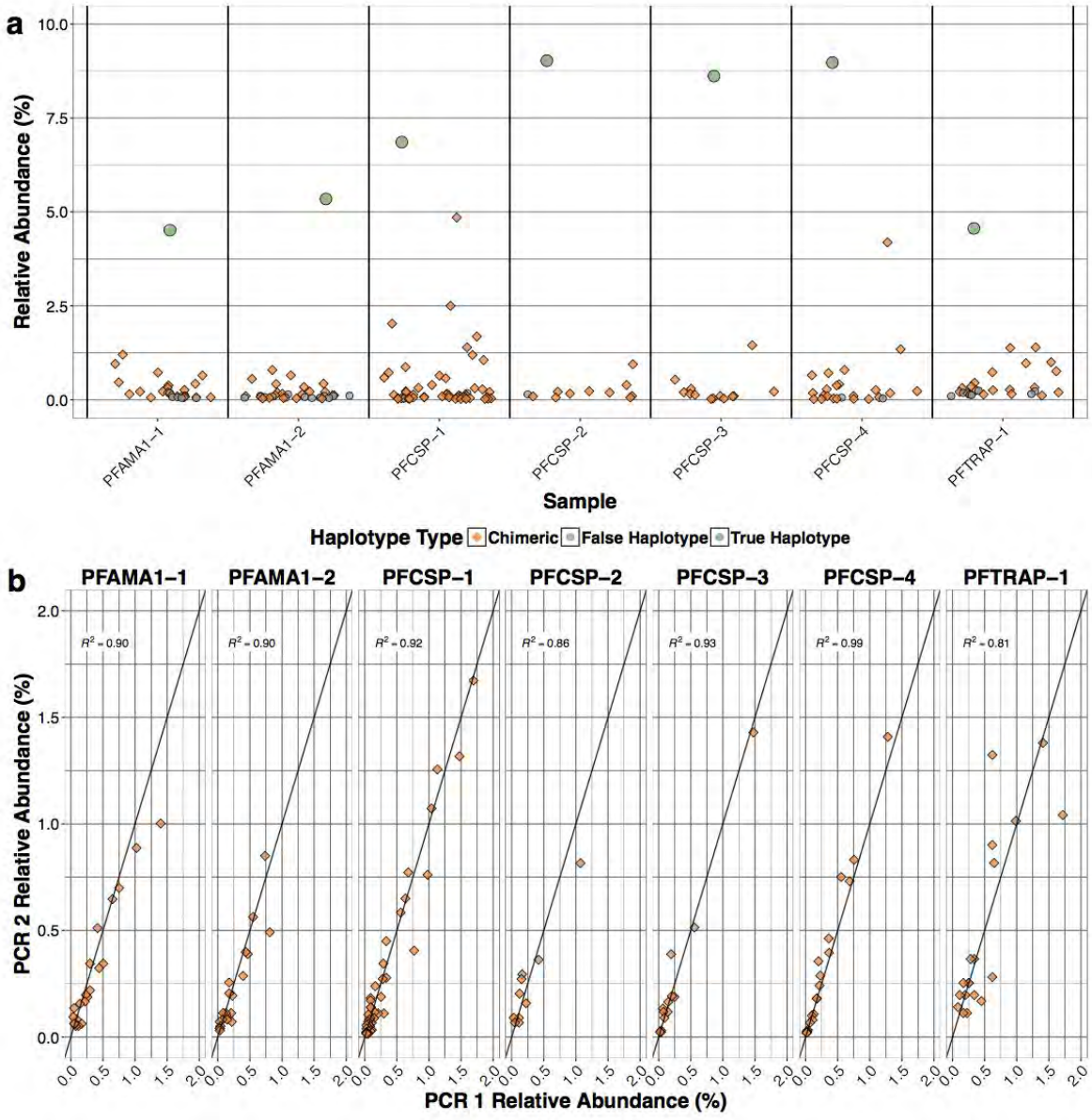


Figure 2.17: Chimera Detection

a) A jitter scatter plot of the SeekDeep results for the haplotypes for samples that had appreciable chimeras with the x-axis being sample and y-axis being predicted relative abundance (truncated at 10%, all haplotype above 10% are true haplotypes). The haplotypes are to appear in both replicates to be plotted. The haplotypes that were marked chimeric are orange diamonds, true haplotypes less than 10% are green circles, and false haplotypes that didn't get marked chimeric are grey circles. **b)** A plot comparing the the predicted relative abundances of the replicates for the false haplotypes which demonstrates the reproducibility of chimera formation across PCR reactions. Deviation from the line of identity represents the difference in the replicates and is summarized by the correlation coefficient.

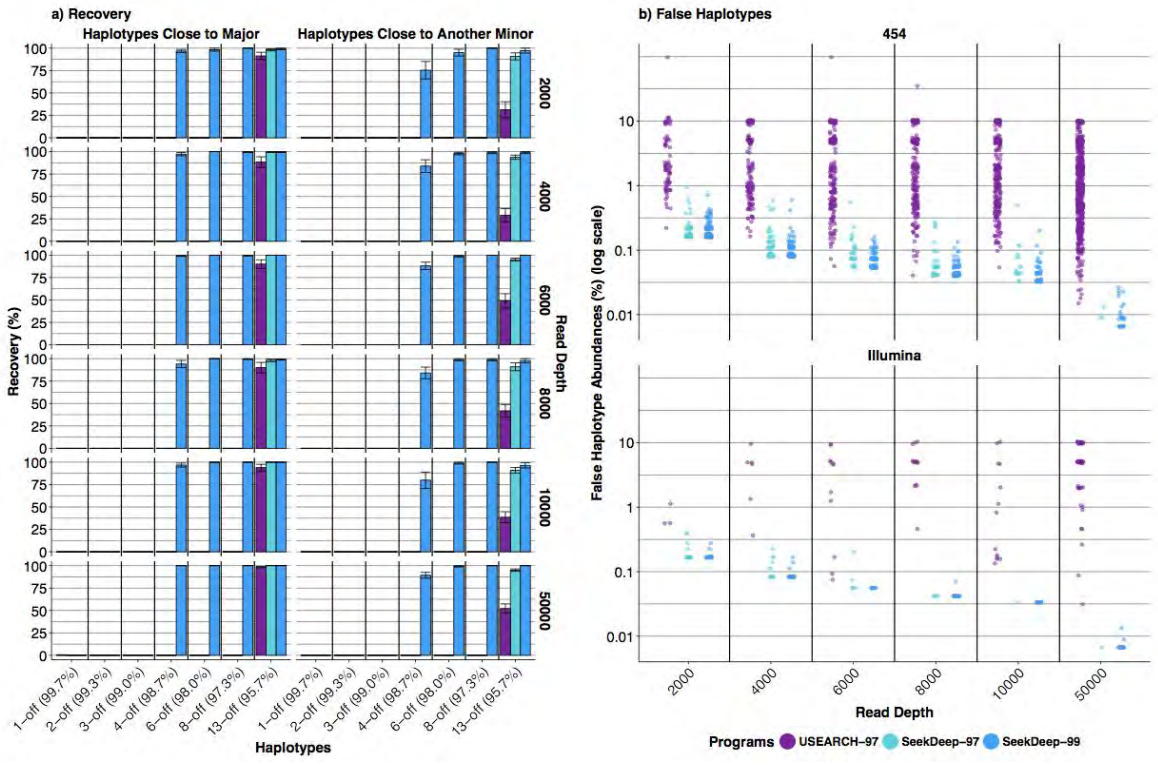


Figure 2.18: OTU Clustering Performance on Simulation Data

SeekDeep offers OTU clustering that is based on only high quality differences rather than any difference. This helps to improve both haplotype recovery and false haplotype creation compared to the OTU clustering offered by USEARCH. Performance of OTU clustering for SeekDeep is shown for both 99% and 97% OTU clustering while only 97% OTU clustering is shown from USEARCH due to a reported bug in the program that only allows 97% clustering. **a)** Haplotype recovery is shown for the two different simulation mixtures depicted in **Figure 2.2**. Haplotype recovery is binned by the degree of difference (and corresponding percent identity) between the haplotypes, and is further stratified by read depth. **b)** A jitter scatter plot is shown of the relative abundance of false haplotypes, stratified by read-depth (x-axis) and sequencing technology. Bars and points are colored by program and OTU level of clustering.

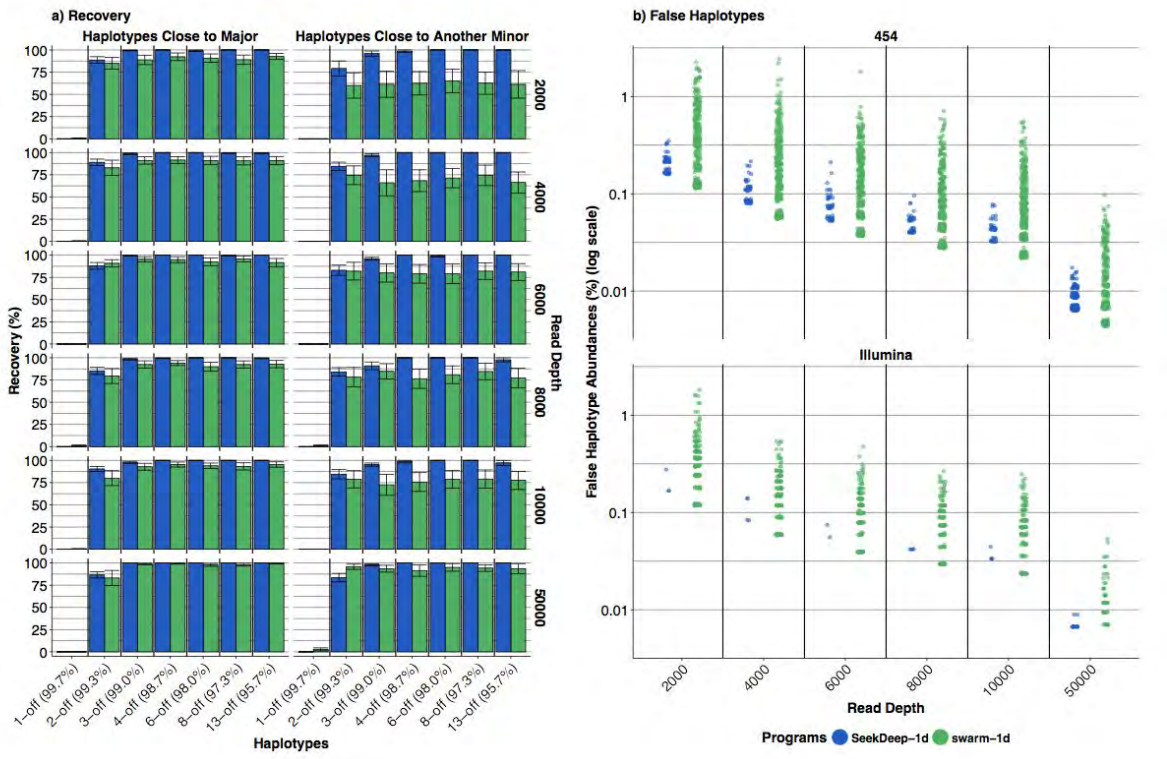


Figure 2.19: Collapsing on Single-base Differences Performance on Simulation Data

SeekDeep, like swarm, can be tuned to account for the number of differences upon which to collapse; however, unlike swarm, SeekDeep can account for type and quality of errors during clustering. Here, we demonstrate the performance of swarm collapsing on 1 difference compared against SeekDeep collapsing on 1 high quality difference and allowing for low abundance and low quality differences as well is shown. **a)** Haplotype recovery is shown for the two different simulation mixtures depicted in **Figure 2.2**. It is binned by the degree of difference (and corresponding percent identity) between the haplotypes, and is further stratified by read depth. **b)** A jitter scatter plot of the relative abundance of false haplotypes, stratified by read-depth (x-axis) and sequencing technology. Bars and points are colored by program.

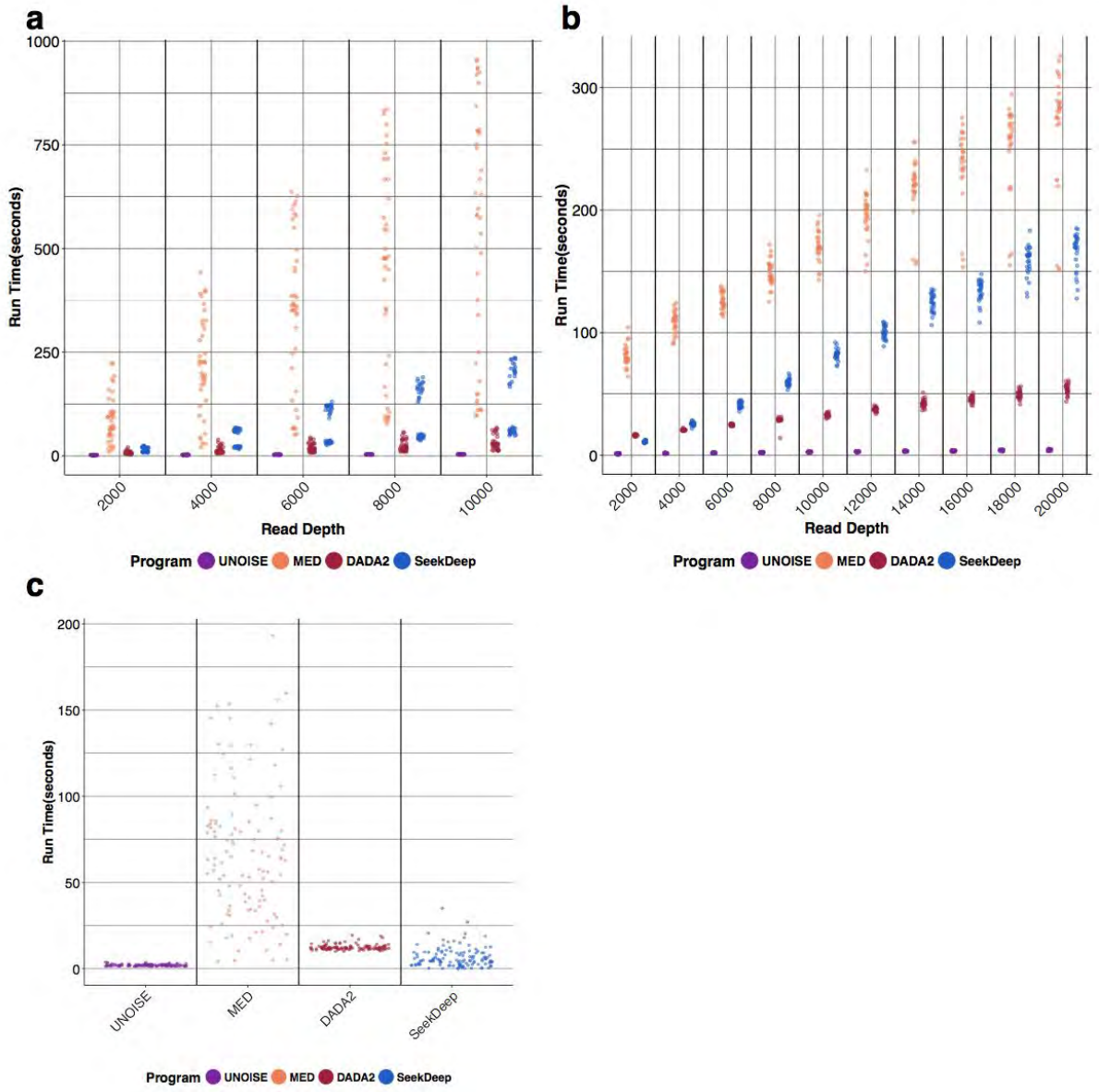


Figure 2.20: Program Run Times

The distribution of program run times in seconds is shown **a)** for simulation datasets across read depths, **b)** for all the randomly down sampled samples from Salipante *et al*. 2014, **c)** for the *in vitro P. falciparum* control datasets. These times should approximate what a user would expect using a personal computer.

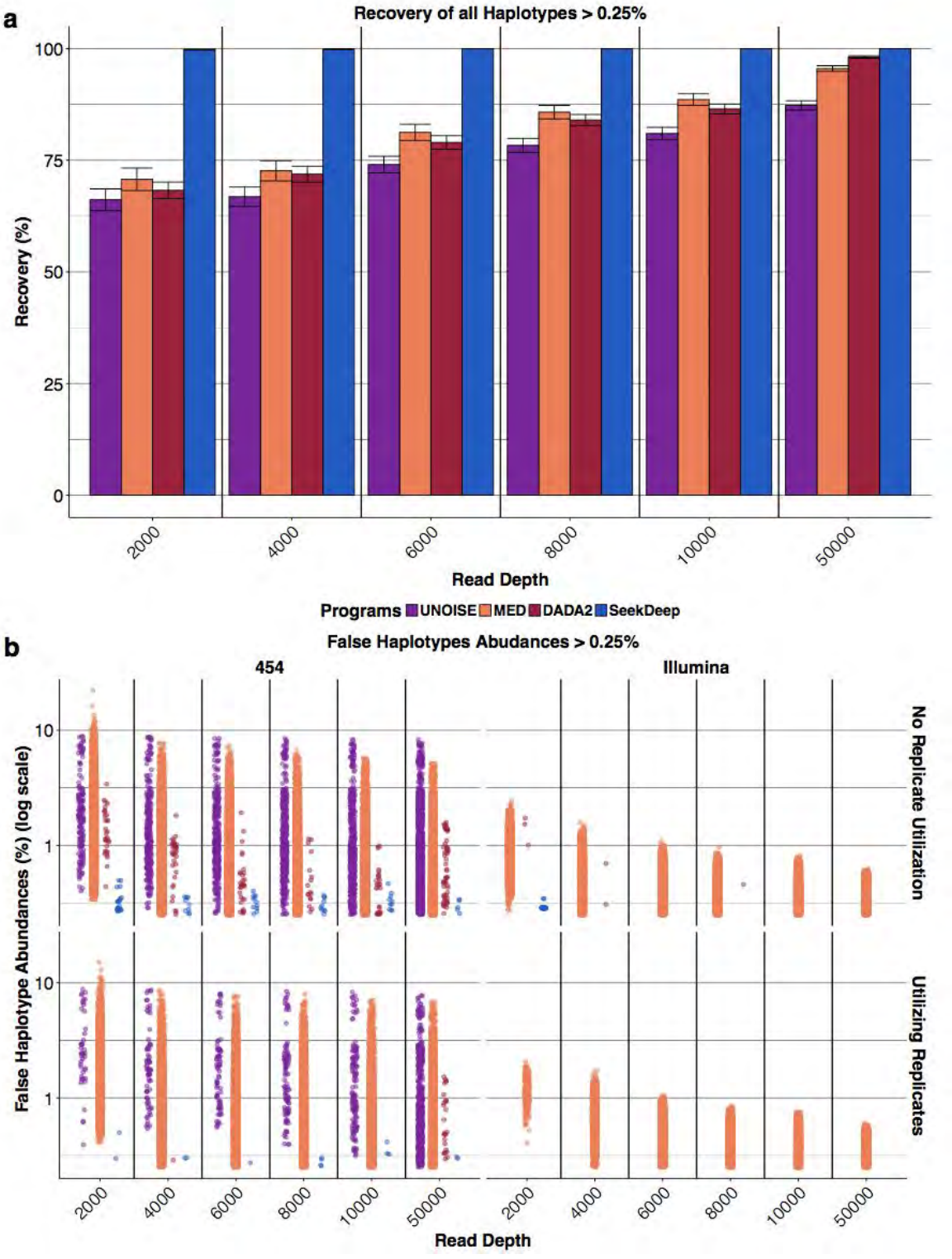

a
Recovery of all Haplotypes > 0.25%
Recovery (%)
100
75
50
25
0
2000
4000
6000
8000
10000
50000
Read Depth
Programs UNOISE MED DADA2 SeekDeep
b
False Haplotypes Abudances > 0.25%
454
Illumina
No Replicate Utilization
Utilizing Replicates
False Haplotype Abundances (%) (log scale)
10
1
Read Depth

Figure 2.21: Haplotype Recovery of Expected Haplotypes and Creation of False Haplotype above ≥0.25% on Simulated Datasets

Performance on the simulations dataset is shown, using a minimum haplotype abundance threshold of 0.25%. **a)** SeekDeep is able to haplotype recovery all expected haplotypes across all read depths simulated. **b)** By making a cut at 0.25% SeekDeep calls practically no false haplotypes. Detection of haplotypes at **≥** 0.25% likely approaches what can be detected by sampling and PCR. At this level of resolution SeekDeep has excellent performance characteristics with better haplotype recovery than other programs with minimal false haplotypes rates that occur only at low abundances.

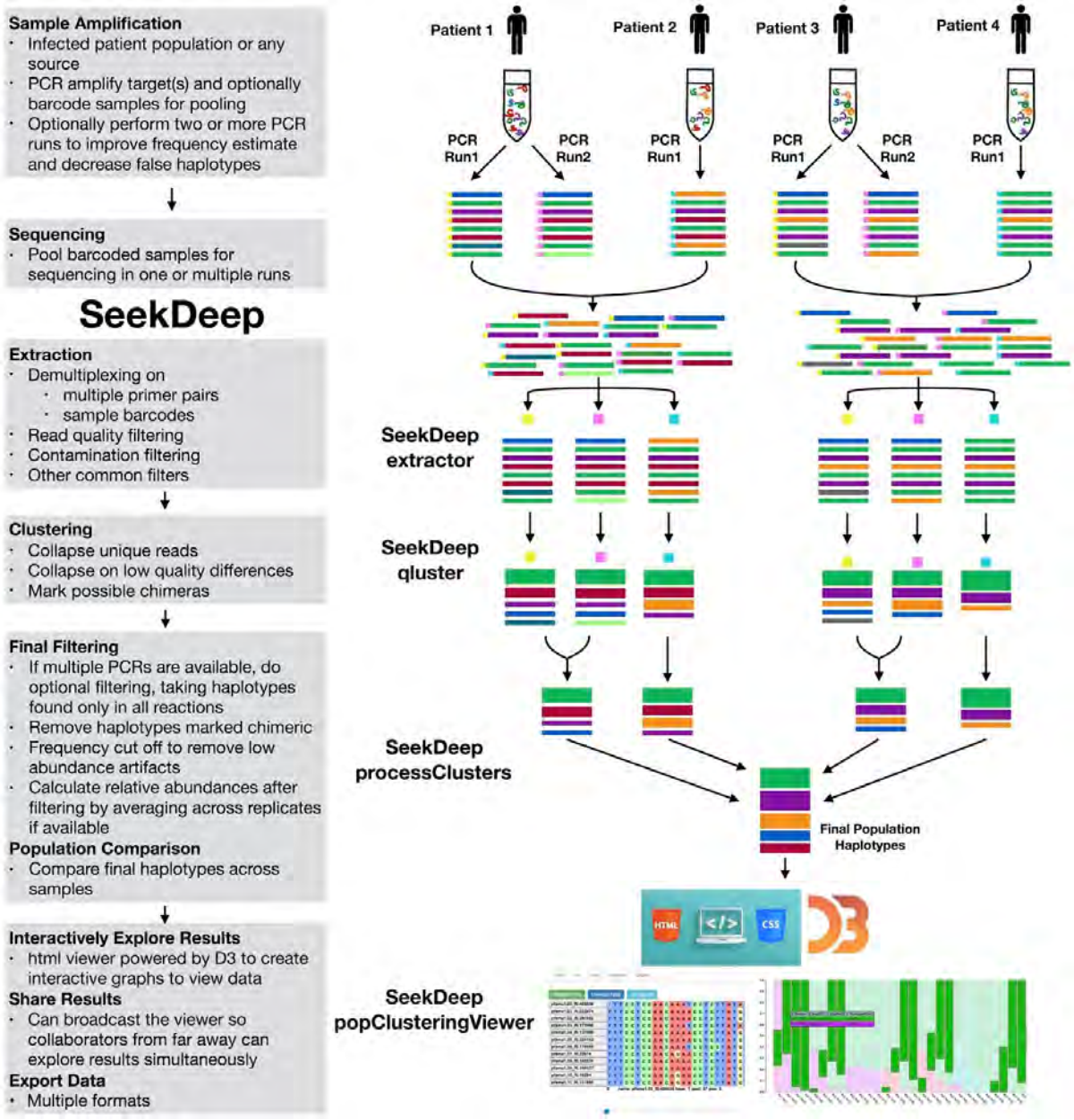


Figure 2.22: SeekDeep Overview

The depicted SeekDeep pipeline was designed to handle diverse experimental and computational workflows. In general, input sequence data is organized as one or more groups of samples that can represent natural populations, different experimental conditions, or any other defined classification. The pipeline is modular, allowing for substitute or additional processing at any step as well as access to the underlying data. The goal of SeekDeep is to perform initial processing and clustering along with exploration of the results and quality control. Extraction is done by **extractor** to demultiplex on sample barcodes (depicted here as colored squares at the beginning of sequences) and/or multiple primers if either are still present in input data. Next, sequences are clustered at the sample level by **qluster** based on either presets for specific sequencing technologies or user defined

parameters to provide the requisite level of resolution (see **Figure 2.23** for how these errors are characterized). Finally the haplotypes generated by **qluster** are analyzed by **processClusters** to take into account replicate comparisons (if available) and then compare sample haplotypes to generate population-level haplotypes and statistics. Final results can be viewed with **popClusteringViewer** in an interactive HTML viewer. For more specific downstream analyses, data can be outputted in multiple formats.

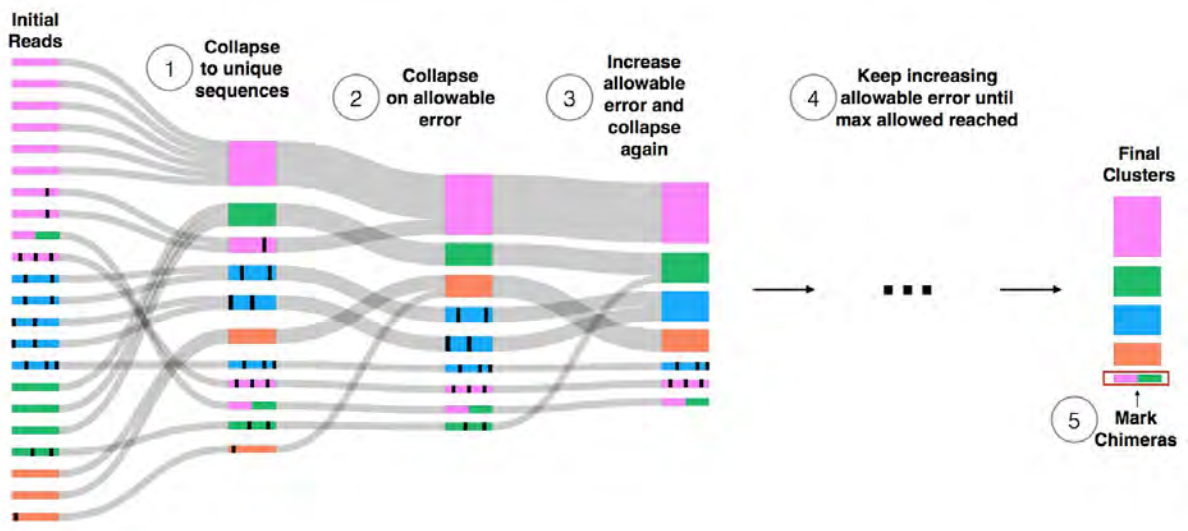


Figure 2.23:Overview of the qluster Algorithm

Qluster starts by operating on the initial reads that have already been demultiplexed by sample and primers. Qluster first **(1)** creates initial clusters of identical sequences and then sorts these clusters by read count in descending order. Pairwise global alignments are then used to compare the representative sequences and **(2)** collapse initially on a minimal amount of allowable errors (see **Figure 2.24** for depiction on how errors are characterized). Qluster thus collapses only the most similar sequences and creating larger aggregate clusters for further comparisons. After each collapse, a consensus sequence is created. On the next iteration comparing all clusters, the amount of **(3)** allowable error is increased to further collapse clusters. Further iterations **(4)** increase amount of error allowed, again creating a consensus after each collapse. Final clusters are created after the final iteration. After clustering, **(5)** mark any sequences that could be a possible chimera.

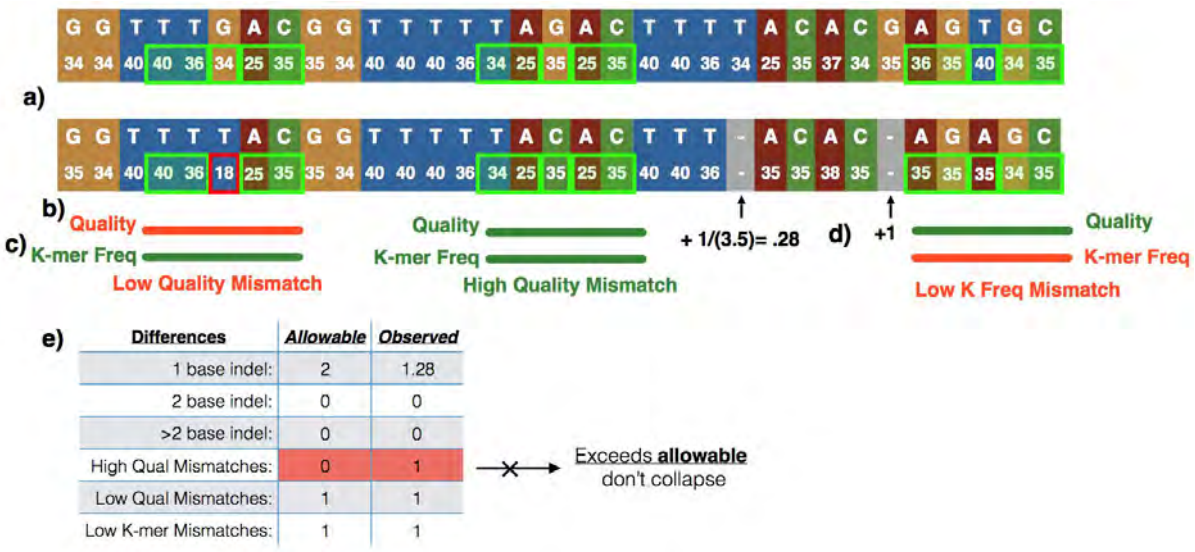


Figure 2.24: Characterizing Errors in Pairwise Comparisons within qluster

Clustering incorporates both base quality and abundance of k-mers as well as parameters relevant to the error profile of the sequencing platform. Depicted is a detailed example of how qluster scores and then determines whether to collapse two clusters. **a)** After pairwise global alignment of the cluster consensus, potential errors are categorized into indels and mismatches. **b)** Mismatches between sequences are first checked for base quality, which includes comparing the base quality scores of mismatching bases and the surrounding bases to a quality thresholds (default is 20 for mismatching bases and 15 for surrounding bases). Both the mismatch site and regional qualities must be higher than this threshold to be considered a high quality mismatch. **c)** High quality mismatches are then further classified by their occurrence in the input data based on the abundance of k-mers in each sequence centered on the mismatch. By default, if the k-mer only occurs once in the input data it is marked as a low abundance mismatch signifying likely error. **d)** Indels are classified by size and are classified into 1-base indels, 2-base indels, and > 2-base indels. Optional weighting for indels that occur in homopolymers can be turned on for pyrosequencing platforms (i.e. 454 and Ion Torrent). **e)** Errors are tabulated and then compared to the current thresholds to determine if the two given clusters should be merged or maintained. In the depicted example, the clusters are not merged as the number of high quality mismatches observed exceeds the threshold for collapse.

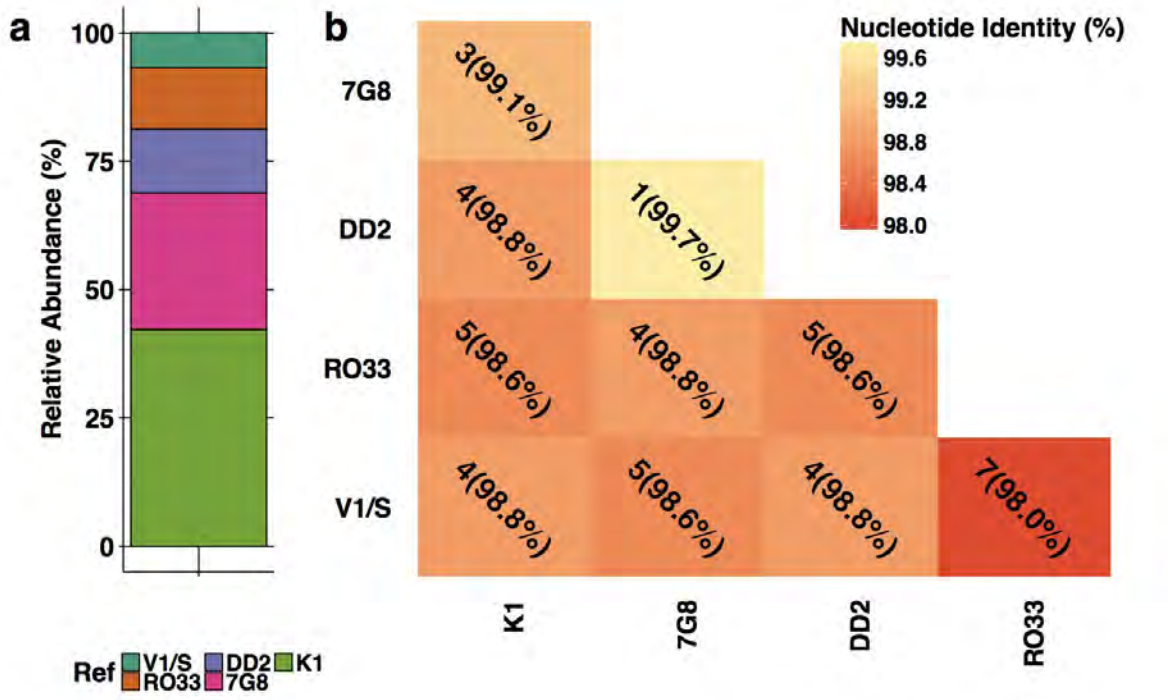


Figure 2.25: *In vitro P. falciparum TRAP* Strain Mixture

The *TRAP* mixture consisted of 5 different *P. falciparum* strains. The mixture was amplified and sequenced twice. Panel **a)** gives the expected relative abundances for the mixture and **b)** is a distance matrix describing the number of base mismatches and percent identity between the strains.

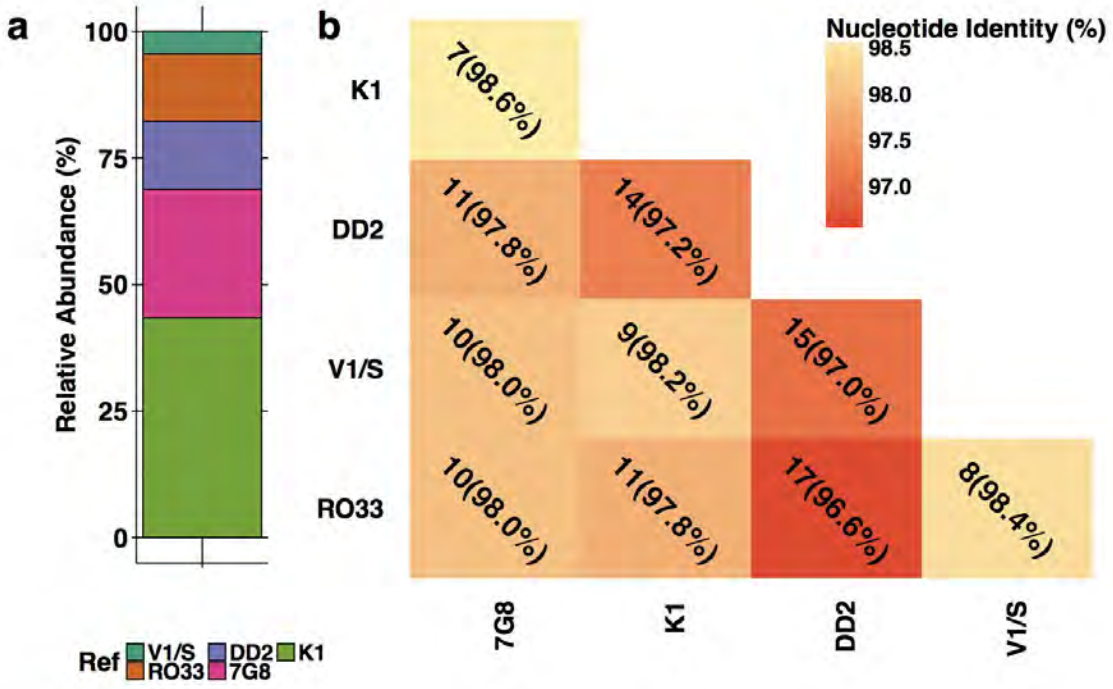


Figure 2.26: *In vitro P. falciparum AMA1* Strain Mixture

The *AMA1* mixture consisted of 5 different *P. falciparum* strains. The mixture was amplified and sequenced 4 times. Panel **a)** gives the expected relative abundances for the mixture and **b)** is a distance matrix describing the number of base mismatches between the strains and the corresponding percent identity.

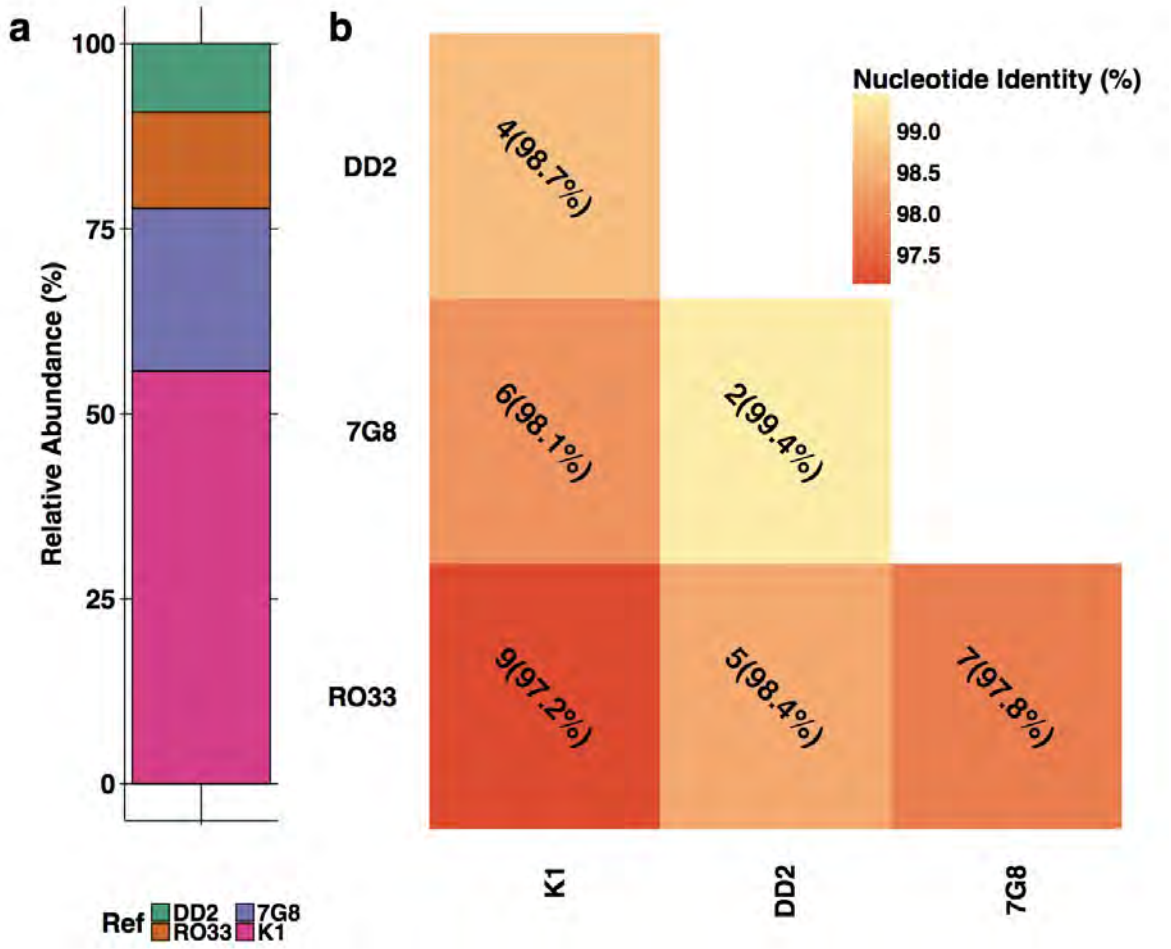


Figure 2.27: *In vitro P. falciparum CSP* Strain Mixture

The *CSP* mixture consisted of 4 *P. falciparum* strains. The mixture was amplified and sequenced 8 times. Panel **a)** gives the expected relative abundances for the mixture and panel **b)** is a distance matrix describing the number of base mismatches between the strains and the corresponding percent identity.

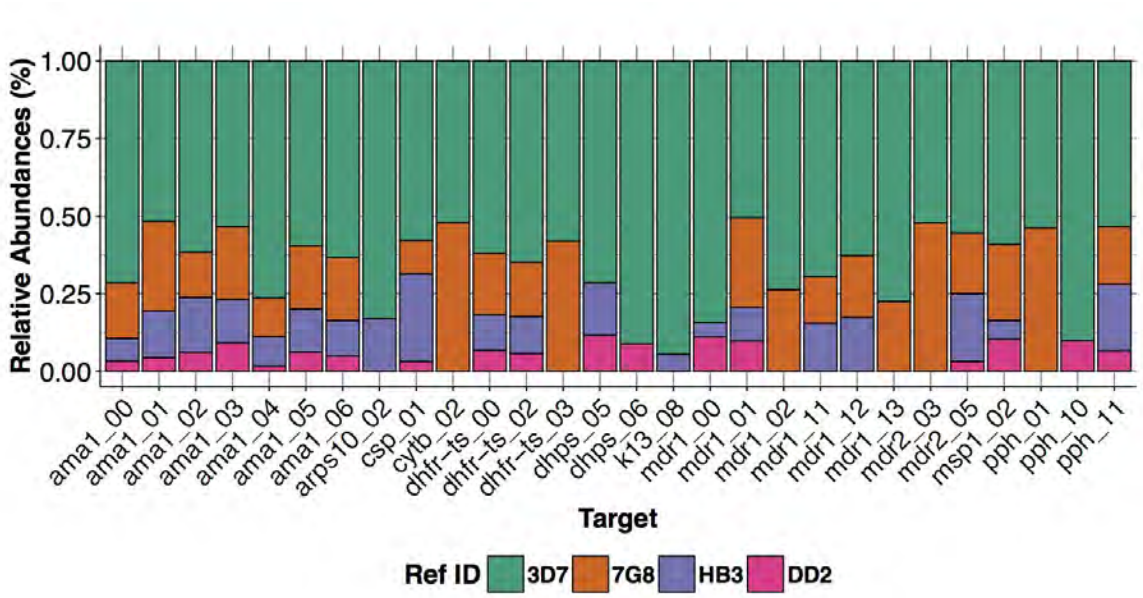


Figure 2.28: *In vitro P. falciparum* Illumina Strain Mixtures

The expected abundances for all amplicons in the control mixture of the strains 3D7, 7G8, HB3, and DD2. While the mixture of the individual strains was constant (3D7 = 79%, 7G8 = HB3 = DD2 = 7%) for all amplicons the strains often shared the same haplotype leading to variation in the number and abundance of haplotypes (2-4) across the amplicons. Differences between strains range from 1-2 SNPs and sometimes large indels (10-15 bp).

Chapter III: kluster: Long Amplicon Clustering using k-mer Similarity Scores

Preface

The following chapter is currently being drafted into a manuscript for submission.

Abstract

Longer amplicon analysis than what is possible with Illumina or Ion Torrent can greatly aid a targeted approach when analyzing highly diverse regions, where shorter target primers cannot be designed, or if the region of interest is longer than the length possible with Illumina and Ion Torrent which could occur when sequencing several SNPs associated with drug resistance that are more than 1kb apart from each other. To that end, PacBio, which has sequence reads at lengths of several kb in length, is the ideal tool; unfortunately, PacBio also suffers from a much higher error rate that hinders analysis. Here, I introduce a novel cluster algorithm that clusters PacBio reads to 1-base resolution, and I test its performance on both *in silico* datasets and known lab strain control mixtures.

Introduction

SeekDeep, which was introduced in the previous chapter, performs well for targeted amplicon analysis when dealing with sequencing from sequencing platforms 454, Ion Torrent, and Illumina; unfortunately, SeekDeep ran into challenges when its pipeline was

tried on PacBio data. Challenges encountered included much higher error rate and the much longer sequencing length which can be several kilobases (kb) for PacBio but only ~400 bp for Ion Torrent, 454, and Illumina. This presents challenges for SeekDeep's pipeline, since the pipeline relies on building solid initial clusters by collapsing on unique sequences and then comparing with global alignments to other sequences. PacBio's high error rate and long sequence length means that the majority of the time no sequences are identical enough to be used to create solid initial clusters; thus the amount of time required to create pairwise global alignments goes up exponentially due to the need to create a matrix of sequence length by sequence length to dynamically determine the best alignment--which means that costly all-by-all pairwise comparison would have to be conducted. Previous attempts have either simply clustered the sequences at an 97% OTU (Schloss et al. 2016) or have attempted to determine local haplotypes by mapping to a reference sequence and correlating SNP variants (Alexander Artyomenko et al. 2016). However, these methods are not adequate if studying regions that can't be mapped to a reference, or if looking for causative single nucleotide differences in drug resistance genes. For these reasons, I developed a novel method for clustering PacBio sequences based off of a similarity score using shared k-mers between sequences which can be used instead of the qluster algorithm described in Chapter II.

The method proves to be much faster than alignment-based comparisons and is sensitive to 1 base pair difference for variants down to 1% abudances. Here, I provide results to validate the method by analyzing mixtures of *P. falciparum* lab strains artificially created in the lab by mixture DNA of known lab strains at specific concentrations. The regions analyzed were a 5kb region of VAR2CSA (a gene that goes under a high level of recombination which prevents mapping to a reference genome), a 1.8kb region of dhfr-ts (a

known drug resistant gene). Previously published control mixtures were also analyzed which included an influenza mixture (Alexander Artyomenko et al. 2016). Also, to determine theoretical limits of detection, a PacBio simulator was created based off monoclonal lab strains and was used to simulated PacBio samples of read depths of 500 to 3,000 reads with sequences of 1.8 kb in length differing by 1, 2, 3, 4, 8,10, and 20 nucleotides.

Results

in silico simulations

All simulated haplotypes differing by 8, 10, and 20 nucleotides showed perfect recovery for all frequencies and read depths. Haplotype recovery for haplotypes that differed by 2, 3, and 4 were 100% recovered at abundances down to 5% but only got up to 80% for lower abundances even at read depths of 3,000 (**Figure 3.1a**). Recovery is dramatically improved (**Figure 3.1b**) when utilizing the method of removing internal clusters based off of SNPs falling 2 standard deviations above the mean error rate (**Figure 3.2**). False haplotypes were rare and never appear above 0.89% abundance and utilizing replicates removed all false haplotypes.

Known Lab Strain Mixtures

An influenza dataset from a previous study (A. Artyomenko et al. 2015) which contained 10 clones with strains and differences ranged from 2 (99.0% identity) to 21 (89.5% identity). All 10 clones were recovered at close to expected frequencies and no false haplotypes were created (**Figure 3.4**).

The *P. falciparum dhfr-ts* dataset consisted of 3 different mixtures of the four lab strains Pf3D7 (major), Pf7G8 (minor), PfDd2 (minor), PfHB3 (minor), with minor strains at 5% (mixture 1), 1% (mixture 2), and 0.2% (mixture 3) abundances (**Figure 3.3**). All strains were recovered for mixtures 1 and 2, but all three minors were missing in mixture 3 (**Figure 3.4**). Recovered strains were close to expected frequencies and no false haplotypes were created.

The *P. falciparum var2csa* dataset consisted of 6 mixtures with 7 different known lab strains at various abundances **(Figure 3.4)** done in duplicate**.** All expected strains were detected in all mixtures very close to expected abundances. One false haplotype was created and this was removed by utilizing duplicates.

Discussion

Longer amplicon analysis can improve the amount of information gained from a targeted amplicon analysis; in some cases, a longer amplicon is needed to encompass the entire region of interest or is needed due to lack of regions with sufficient sequence conservation to design shorter target primers. PacBio can generate sequences of several kb in length, but also suffers from a high error among other issues. In order to take advantage of the longer read lengths offered by PacBio but still have single base resolution, the kluster algorithm was added to the SeekDeep pipeline described in Chapter II. The kluster algorithm works by creating connections between sequences based on similarity scores created by counting the number of k-mer shared between sequences at various k-mer lengths. In this way, a connected graph of reads is created and clusters determined with a density-based spatial clustering of applications with noise (DBSCAN) approach to avoid over-clustering.

Kluster is able to detect single base differences in sequences down to 10% abundance at read depths of 500-3,000, but starts to over-collapse the closely related strains below that (**Figure 3.1**). To further help detect closely related strains at low frequencies that may have been over-collapsed, reads within a cluster that all contain a variant detected at a frequency higher than expected based on the error rate calculated across all clusters are removed to form their own cluster; this dramatically increases haplotype recovery without creating more false haplotypes (**Figure 3.1**).

Using kluster, we have shown that it has perfect recall of a mixture of 4 *P. falciparum* strains all related to another strain by one difference on a region of Pfdhfr-ts, an important gene involved in drug resistance in *P. falciparum*. In *P. falciparum* even a single difference in a gene can lead to drug resistance, and 1-base resolution is paramount when sequencing such genes. Here, we have demonstrated that the novel algorithm, kluster, is able to cluster sequences with single base resolution even at low read depths.

Methods

Datasets

in silico simulations

PacBio Simulator

At the time of writing, there are no available simulators for PacBio that will do targeted amplicon sequencing, so an in-house simulator was created. This was done by utilizing 10 monoclonal FCR3 *var2csa* PacBio samples and 20 monoclonal 3D7 *var2csa*

PacBio samples and first aligning the samples to their expected sequences. The PacBio technology currently reports quality values up to a max of 42, and the per base error rate was calculated for bases with quality of 42 and assumed to be the PCR error rate. This theoretical PCR error rate was subtracted from overall error rate to get a PacBio error rate per base. Quality score distributions were then created for both mismatches and matches after subtracting the theoretical PCR error rate for each quality score. Insertions and deletions of up to 5 bases were observed, and the per base indel rate was calculated; a size distribution for both insertions and deletions was created based off of the counts observed. No correlation between position and error rate was observed (unlike Illumina and Ion Torrent that show positional effects). This is not unexpected due to the circular nature of PacBio sequencing and the fact each position gets several sequencing passes rather than just being sequenced once in a linear fashion; the same was true for rate of insertions and deletions.

Using the calculated rates and distributions, a simulator was written to take a sequence and simulate per base whether there was no error, a mismatch, a deletion, or insertions. If a match, the quality score is determined by pulling from the match quality score distribution. For mismatches, the mutated base was based off the observed substitutions rates per base, above which favored transitions over transversion and the quality score was pulled from the mismatch quality score distribution. The size of a simulated deletion or insertion was determined from the appropriate size distribution; if an insertion was simulated, the bases inserted was randomly generated using the base composition of the input sequence. In this way, a simulator was created that could take an input sequence and emulate reads that would result from PacBio sequencing. This was combined with the

simulator described in Chapter II, which simulates PCR by taking PCR cules into account, with errors occuring in earlier cycles appearing at higher abundances.

Simulated Datasets

Datasets representing several different abundances, read depths, and sequence identities were simulated using the above simulator to test the theoretical bounds of kluster. Each dataset consisted of 7 minor strains and 1 major strain for a total of 8 strains per mixture. The minor strains differed from the major strain by 1, 2, 3, 4, 8, 10, and 20 SNPs, and did not share SNPs. A template haplotype of *P. falciparum* 3D7 *ama1* was used as the major strain and SNPs were randomly generated off this template to generate the minor strains. Mixtures were simulated with all the minor strains at the same abundance, with the major strain taking up the rest of the mixture. The minor strains abundances were 10%, 5%, 2%, 1%, and 0.5%. Each of these abundance datasets were simulated twice to emulate duplicates and were simulated at read depths of 500, 1,000, 1,500, 2,000, 2,500, and 3,000. This was done 10 times, with new SNPs generated each time. This resulted in 600 simulated datasets; kluster was evaluated for its ability to recover all expected sequences, and this was averaged across the ten different sets of randomly generated SNPs.

Influenza

An influenza dataset from a previous study was also analyzed (A. Artyomenko et al. 2015). The amplicon was 2kb and was a mixture of 10 clones at relative frequencies of 50%, 25%, 12.5%, 6.25%, 3.125%, 1.56%, 0.78%, 0.39%, 0.19%. The clones were closely related and differences ranged from 2 (99.0% identity) to 21 (89.5% identity) and sequenced at a depth of 18,134.

Plasmodium falciparum

dhfr-ts

A 2kb region of the *dhfr-ts* gene of *P. falciparum* was amplified in three different mixtures of the lab strains 3D7, HB3, 7G8, Dd2 in duplicate. The three mixture were HB3, 7G8, and Dd2 all at 5% (mixture 1), 1% (mixture 2), and 0.2% (mixture 3) with the rest of the mixture being 3D7 for each mixture. For the *dhfr-ts* region each strain differed from at least one other strain by 1 mismatch while the differences between strains for the ama1 region ranged from 16 (99.2 identity shared) to 28 (98.5% identity). Read depth for samples ranged from 557 to 2800.

var2csa

Six different mixtures of known lab strains of *P. falciparum* that were amplified in duplicate for a 3kb region of the *var2csa* gene were analyzed. Strains were very distantly related to each other (90.6% to 92.2%) and sample read depths ranged from 119 to 319.

Algorithm Overview

K-mer Similarity Score

First, I define a k-mer similarity score for a given k-mer length of k as the total number of shared k-mers between the two sequences divided by the total number of possible k-mers shared (which is the length of the shorter sequence minus k plus 1). A score of 0 would mean that there are no k-mers shared between the sequences, and a score of 1

would mean that all the k-mers of the shorter sequence can be found within the longer sequence.

Graph Based Clustering

The algorithm starts with calculating all pairwise comparisons with k-mer similarity scores which has the option to be parallelized if multiple CPUs are available. A graph is created, with nodes as sequences and the edges connecting the nodes are undirected with weights as the similarity scores.

The goal of the first round of clustering is not to cluster all sequences that belong to one haplotype into 1 cluster: rather, the goal is to gather together enough sequences into each cluster to ensure that when a consensus sequence is created from this cluster, it creates the correct consensus for the local haplotype it belongs to--thus, clusters that end up creating the same consensus can then be further clustered together (**Figure 3.5**). This allows the initial clustering to be strict enough to minimize clustering together similar haplotypes. Several attempts at optimizing k-mer length and a k-mer similarity score cut off for making connections in the graph were attempted but it was found that each dataset with different read lengths had different optimal k-mer lengths and k-mer similarities that were able to recover all expected sequences. Therefore, a new score was calculated to make connections: first, calculate the k-mer similarity scores between sequences for k-mer lengths of 2, 3, 4 and 5 and take the slope in k-mer similarity between lengths to calculate a distance score to be used in edge connections. This approach was settled upon from out of the several approaches and different scores attempted because it proved to be able to recover all expected sequences for different read lengths and different species datasets; the approach could also be calculated quickly, as the time it takes to calculate scores for these

k-mer lengths is small compared to longer k-mers. Clusters are then created by using a density-based spatial clustering of applications with noise (DBSCAN) approach (Ester et al. 1996). In short the DBSCAN algorithm works by taking a node connecting all nodes that are connected under a certain distance, called epsilon, and if the number of nodes connected is greater than or equal to a set number, called minimum number of neighbors, then the nodes are clustered into one group. Once neighbors from a single node are connected, one of these neighbors is chosen and its neighbors are connected under the epsilon. Nodes that do not meet the minimum number of neighbors requirement do not spread to their neighbors and are considered edge points while nodes that do have the minimum neighbors and do spread are considered center points. Nodes that are not center or edge points are considered noise points and do not fall into any clusters. Nodes are chosen at random and classified until all nodes are classified as either center, edge, or noise points and interconnected points are considered a cluster. The clustering for sequences are carried out with a default epsilon of 1, chosen based off simulations where only sequences originating from the same original haplotypes had a slope of decreasing k-mer similarity below 1, and a minimum number of neighbors of a default of 4.

Once final clusters have been created the original raw PacBio sequences are then mapped to the final consensus sequences to determine final read count for each cluster. Sequences that differ by more than a certain percent identity (default 90%) are placed in a separate file which can be investigated for possible missed haplotypes. To avoid clustering together very similar sequences, a per base error rate is calculated from the remapped sequences to the final consensus and then on a per consensus sequence basis, sequences are removed from a cluster to form their own cluster if they all share the same differences to the consensus sequence and if that difference is at least 2 standard deviations (SD) from

the mean per base error rate (**Figure 3.2**). Final consensus sequences are given in a fastq file with quality scores being the average quality score for that base along with the number of reads for each given cluster (**Figure 3.6**). These final results can then be given to SeekDeep's processClusters function to determine shared haplotypes between samples.

Figures

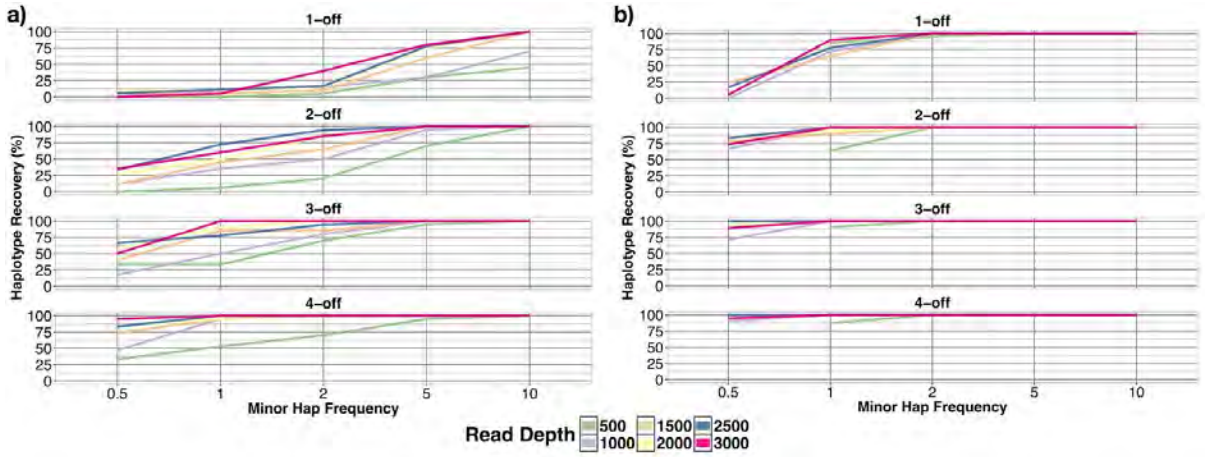


Figure 3.1: *in silico* Haplotype Recovery Results

Haplotype recovery for the *in silico* datasets. This was calculated as the number of times the haplotype was recovered divided by the total number expected which was 10. The x-axis is the abundance for the minor haplotypes, the lines are colored by read depth, and the plots are paneled where each panel is one minor haplotype and the title indicates the number of differences from the major haplotype it is. **a)** is haplotype recovery without removing sub clusterings that contain the same SNPs that fall two SD above the mean error rate observed and **b)** is haplotype recovery when this feature is utilized which greatly improves the recall for all haplotypes down to 1%.

Figure 3.2: Removing Internal Clusters on Shared SNPs

A feature is offered to attempt to recover closely related haplotypes that were improperly clustered with another haplotype where the per base error rate is calculated across all clusters. The PacBio error rate is randomly distributed across the reads and so it's unexpected for the same error to occur in the same base positions multiple times. Therefore, any SNPs that occur at a rate greater than 2 standard deviations above the global error rate are determined and reads containing that SNP are removed to form their own cluster.

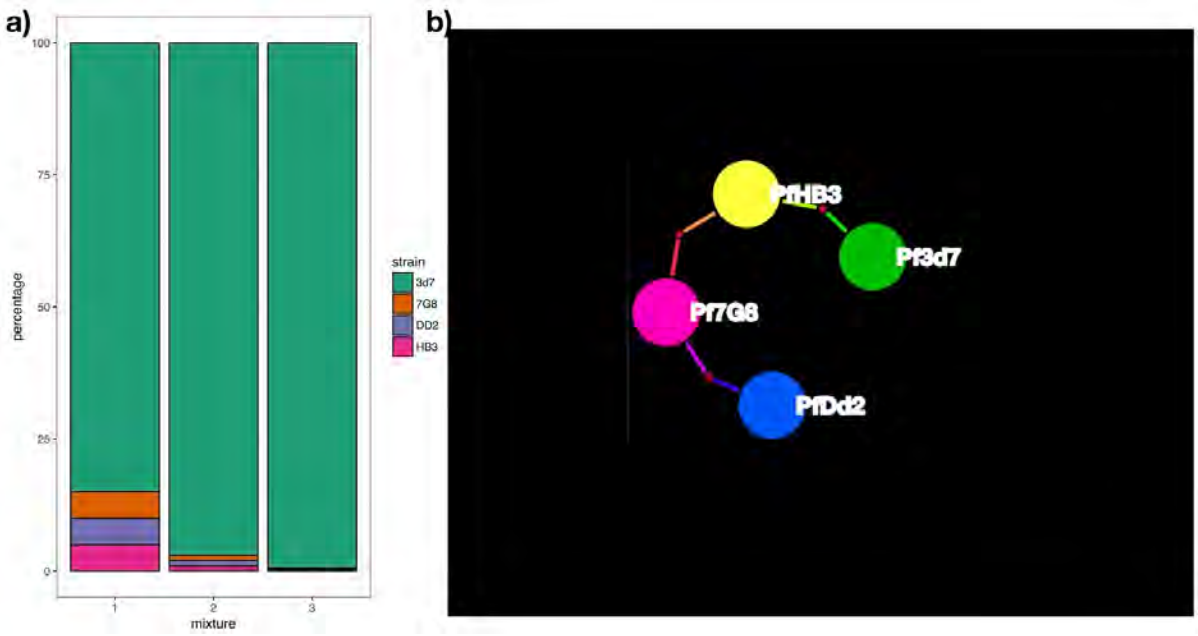


Figure 3.3: *P. falciparum dhfr-ts* Mixture Setup

a) The relative frequencies of the lab strains in the three mixtures. **b)** Each strain is one base different from at least one other strain, the number of red dots on the lines connecting strains is the number of difference between them.

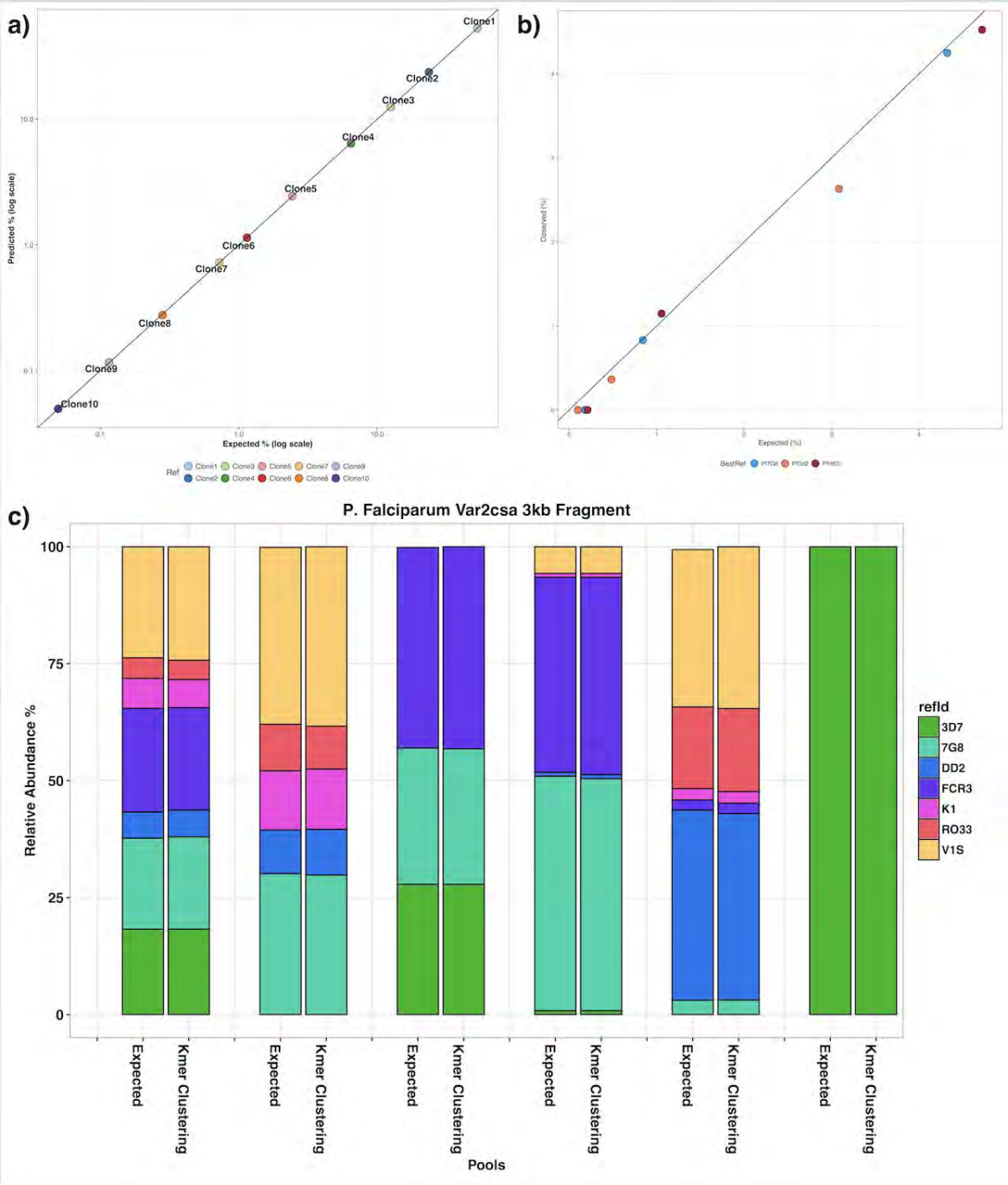


Figure 3.4: kluster Results on Known Lab Strain Control Mixtures

a) kluster recovers all ten of the influenza clones and created no false haplotypes, x-axis is expected abundance and y-axis is kluster's abundance for the clone, black line is line of identity. **b)** kluster recovers all 3 minor strains at 5% and 1% mixtures but fails to recover all at the 0.1%, no false haplotypes were created, x-axis is expected abundance and y-axis is

kluster's abundance for the clone, black line is line of identity. **c)** kluster recovers all strains in all mixtures for *P. falciparum var2csa* datasets, the first bar represents the expected the frequencies and the second bar is the frequencies obtained from kluster which matches very close to expected.

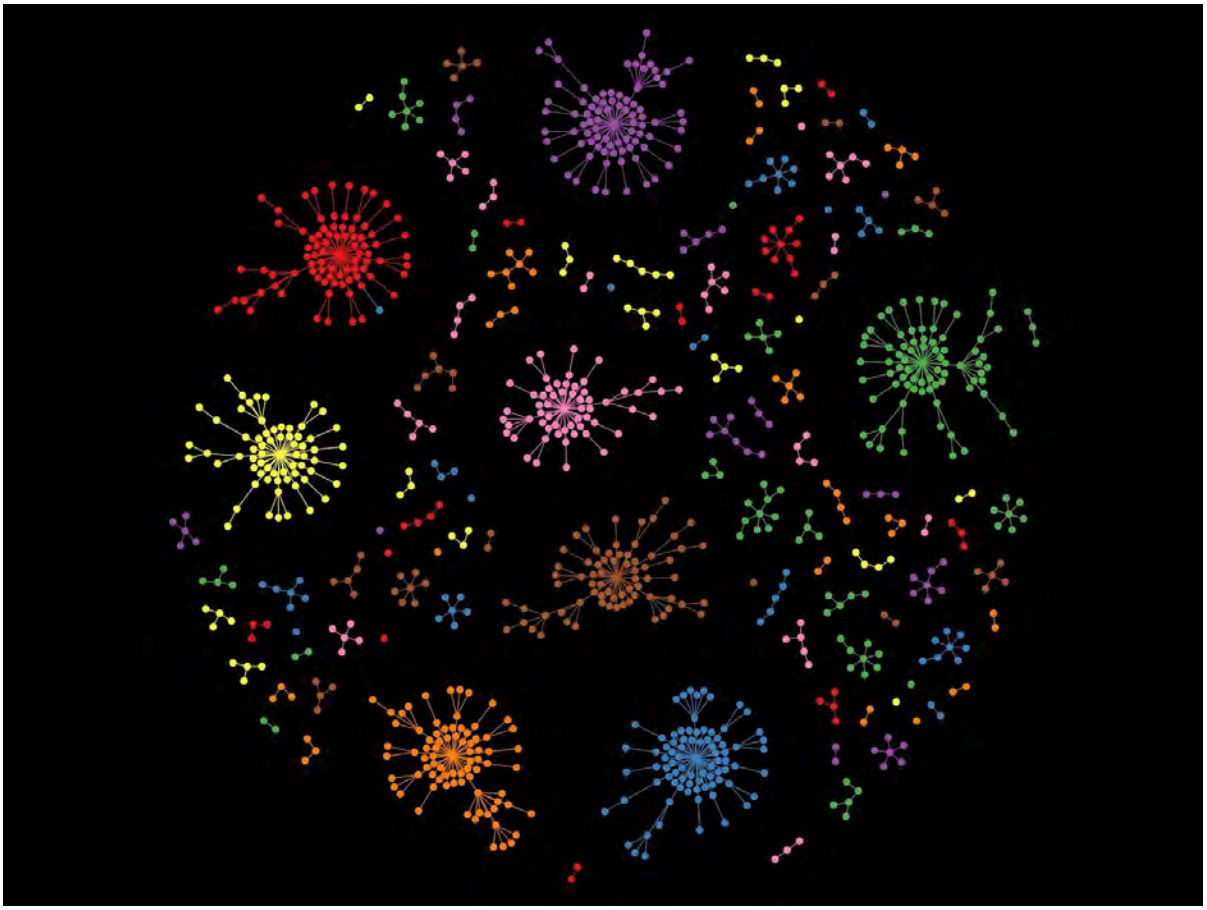


Figure 3.5: Example Initial Clustering Step

This is an example of the first round of cluster described in the methods section. The nodes are colored by the strain they belong to. Multiple small clusters of the same color can be seen and that is because the goal of the first step of clustering is not to gather every single read that belongs to the same strain but rather to gather enough reads to gather that when a consensus sequence is created for each cluster and it will match the consensus sequence of clusters coming from the strain they belong to.

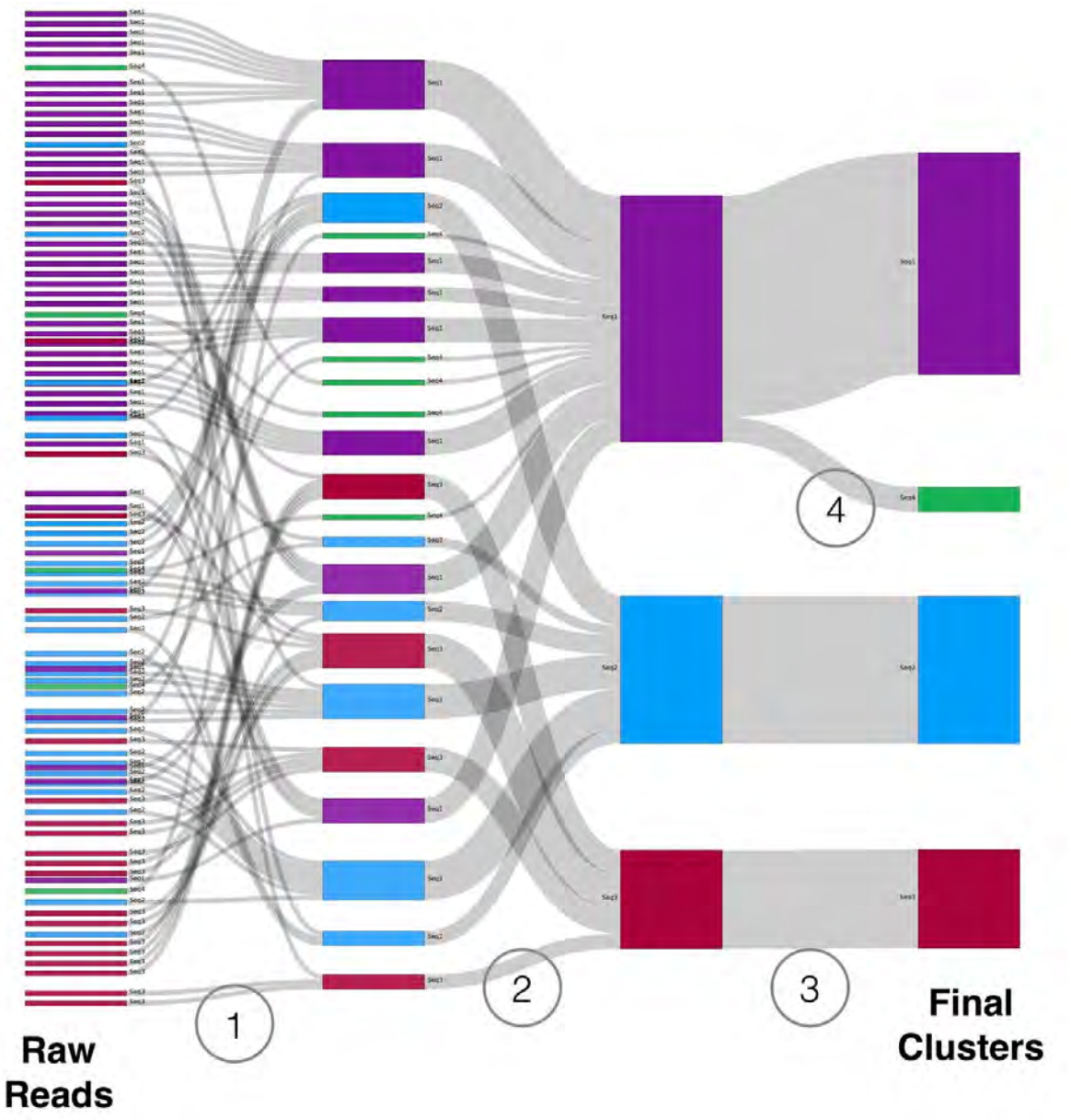


Figure 3.6: Workflow Overview

1) Initial reads are taken and clustered based on k-mer similarity scores to create initial clusters. **2)** Consensus are created for these clusters and each consensus is compared to collapse clusters with the same consensus to create clusters again. **3)** final clusters are used to map the raw reads to better determine read abundance. 4) optionally clusters are checked internally to remove any reads that all contain SNPs that appear at a higher abundance than expected when comparing to a base error rate calculated across all clusters by taking internal reads and comparing to consensus sequence.

Chapter IV: Global antigenic diversity and copy number polymorphism of *var2csa* the leading vaccine candidate for placenta malaria

Preface

The following is adapted from a manuscript being prepared for submission.

Abstract

Pregnant woman can be infected with the *Plasmodium falciparum* species expressing VAR2CSA protein which primarily binds to placental chondroitin sulfate (CSA), leading to sequestration of parasites in the placenta and poor birth outcomes. Antibodies against VAR2CSA has been found to be protective in multigravid women and for this reason the minimum CSA binding ID1-DBL2x-ID2a has been used in two vaccine trials; however, the trials might be hampered by the high degree of diversity of VAR2CSA. For this reason, we have developed a novel program PathWeaver to extract VAR2CSA sequence from publicly available shotgun whole genome sequenced field samples to better characterize this diversity globally and across time. We have found 4 major and 2 minor groups within the ID1-DBL2x-ID2a region that are stable across time and space; this stability is suggestive of balancing selection as well as evidence confirming previous reports of possible VAR2CSA copy number variation.

Introduction

The protozoan disease malaria is still endemic in much of the developing world, infecting an estimated 216 million people per year and causing 445,000 deaths in 2016 (WHO 2017). Pregnant women are particularly susceptible to malaria; *Plasmodium falciparum*-infected erythrocytes sequester in the placenta and can cause poor birth outcomes (Rogerson et al. 2007; Salanti et al. 2003b; Tuikue Ndam et al. 2005). Placental sequestration is mediated by a highly variable protein called VAR2CSA, that primarily binds to placental chondroitin sulfate (CSA) (Rogerson et al. 2007; Duffy et al. 2006; Salanti et al. 2004). Naturally-acquired antibodies to VAR2CSA have been shown to be protective during pregnancy (Rogerson et al. 2007; Ataíde, Mayor, and Rogerson 2014). Efforts to develop a VAR2CSA vaccine are underway (Fried and Duffy 2015; Tuikue-Ndam and Deloron 2015). However, their efficacy may be hampered by the genetic and geographical variation in the protein.

The gene *var2csa*, like most *P. falciparum var* genes, has two exons and is composed of multiple Duffy binding-like (DBL) domains along with a transmembrane domain. The first exon contains 3 DBLX domains and three DBLε, with the 6th and last domain traversing into exon 2. Most *var* genes can be classified by their 5' upstream (ups) region and fall primarily into UpsA, UpsB, UpsB/C or UpsC groups, but VAR2CSA is the only *var* with that has UpsE. UpsE also contains an ups open reading frame (uORF) that encodes 119 amino acids that ends 274 base pairs before the VAR2CSA start codon. This uORF was confirmed to be part of the mRNA transcript of the VAR2CSA transcript although it's not clear if it's translated or not (Lavstsen et al. 2003).

One particular region of VAR2CSA, ID1-DBL2x-ID2a, appears to be most responsible for placental cytoadherence and will hereafter be referred to as the minimum CSA binding domain (MCBD) (Srivastava et al. 2011; Clausen et al. 2012). Antibodies directed to this minimal binding region have shown to occur naturally in woman protected from malaria and have been shown to block binding(Bigey et al. 2011; Salanti et al. 2010). This region is the target of two current vaccines undergoing clinical trial and vaccine efforts of several groups has recently been summarized (Chêne et al. 2016).

In a study of pregnant women in Benin and Malawi, we found that the gene is highly variable with 152 variants in 101 clinical malaria isolates. Previous studies have found two regions within the MCBD to be dimorphic, one region in the beginning of the DBL2 region (VAR2CSA3D7 amino acids 589-617) (Sander et al. 2009) and the other region takes up the majority of ID1 region (VAR2CSA 3D7 amino acids 397-568) (Doritchamou et al. 2015). The combinations of these two regions creates 4 subtypes with one of the types being found exclusively in multigravid women (Doritchamou et al. 2015). The global distribution of these types has yet to be adequately described and it is not known if all types are found globally.

Beyond polymorphisms, copy number variation of var2csa has also been recognized (Sander et al. 2011). The lab strain HB3 has two copies of VAR2CSA although it is unclear if this was a culture adaptation or was naturally present in the strain before culturing. Unlike other PfEMP1 vars where only one copy at a time is expressed, these copies showed coexpression in two field isolated and HB3 with confirmed two copies of *var2csa* (Sander et al. 2009). The genomic positions of these copies were estimated using using pulsed field gel separation of chromosomes and placed HB3's on chromosome 1, confirmed with assembly, and the position of the two field isolates' copies on chromosome 8 and somewhere on 5-8 (Sander et al. 2009). Further investigation of 111 natural isolates from Sudan and Tanzania

showed that 20% of isolates had multiple copies *var2csa* (Sander et al. 2009). Follow up study of isolates from the Cameroon found frequent multicopy isolates as well and determined that multiple copies were associated with pregnancy and increasing gravida. Infections with multicopy strains had longer persistence further suggesting a survival advantage for placental parasites carrying multiple copies (Sander et al. 2011). While such studies suggest var2csa copy number polymorphism may be common occurrence, a full survey of the extent of copy number across the world is lacking.

Recently, thousands of *Plasmodium falciparum* genomes have been shotgun whole genome sequenced with primarily 100-base paired end Illumina sequencing(https://www.malariagen.net/projects/pf3k, Plasmodium 100 Genomes initiative, Broad Institute (https://www.broadinstitute.org)). However, because of the high variability of *va2csa* and its proximity to the telomere, it has been difficult to assess the gene using standard read mapping to reference genome or assembly approaches. Previous studies have tried to use standard de novo assembly programs to analyze complex *P. falciparum* genes (Crosnier et al. 2016; Jespersen et al. 2016; Dara, Drábek, et al. 2017) but these were meant for monoclonal sample assembly and were not built to handle mixtures of multiple genomes (polyclonal samples) or samples with increased gene copy number. This can lead to chimeric assembled sequences where sequence from one strain or copy are falsely combined with sequences from another copy of the gene to create false sequence.

Here we present PathWeaver, a new method which leverages initial read recruitment to a region of interest in a reference genome followed by iterative de novo local assembly and recruitment of unmapped reads in order to assemble highly variable genes that are not amenable to analysis using standard short-read reference mapping methods. Using publicly available whole genome sequencing data, we use this method to interrogate *var2csa*

diversity and copy number variation in order to comprehensively describe genetic variation, genomic signatures of selection and global population structure. These data provide critical information about the potential impacts var2csa diversity may play on the successful development of PAM vaccines.

Results

Assembly on VAR2CSA Upstream Region and Exon 1 (UpsE-ID5)

The PathWeaver algorithm, described in the methods section, was run on all datasets. Subsequent analysis was then performed on full length contigs and then subsequently on contigs that spanned smaller subregions of interest. These subregions were the five DBL regions, their inter domains, the minimal CSA binding domain (Bordbar et al. 2012), and the previously found dimorphic regions in ID1 (Doritchamou et al. 2015) and in DBL2 (Sander et al. 2009).

Mapping Characteristics of in silico Simulated Sequences

The 30 unique *var2csa* UpsE-ID5 sequences collected from the Pf3k Pacbio genome assemblies and from previous studies (Rask et al. 2010) were used to *in silico* simulate shotgun 2x100 Illumina sequencing runs with approximately 40 per base read coverage using a custom shotgun simulator and an Illumina simulator (Huang et al. 2012). On average 90.81% of the simulated sequences aligned to 3D7 (range 81.66%-97.48%), with on average 95.41% of the paired sequences both mapped together (range 92.48%-98.08%) and the rest of the pairs only had one mate mapped. Sequences mapped to *var2csa* 98.93% (range 96.24%-99.95%) of the time, while an average of 1.2% sequence reads mapped to

other *var* genes (range 0.07%-3.38%) and an average of 0.28% mapped to other genomic locations (range 0.07%-0.65%). Mapped sequences were then examined for the extent of soft clipping, with on average 31.57 bases being soft clipped (range 28.08-35.73). This suggests that like most *var* genes, *var2csa* cannot be assembled using reference based variant calling pipelines and thus necessitate novel assembly approaches.

Performance on in silico Simulations and Monoclonal Lab Strains

In order to show that PathWeaver accurately recruits reads and reconstructs the UpsE-ID5 region, the *in silico* simulation datasets were analyzed. In all cases, PathWeaver assembled a single contig which perfectly matched the expected sequence. Shotgun short reads of monoclonal samples from publicly available laboratory strains, for which the *var2csa* sequence is known, DD2, GB4, IT/FCR3, W2, 7G8, and 3D7 were also analyzed by PathWeaver. Each produced a single contig, perfectly matching the expected *var2csa* sequence. This suggests that other *var* genes aren't being erroneously recruited and assembled.

Performance on laboratory strain mixtures (Pf3k Controls)

Twenty eight lab control mixtures generated by MalariaGen were then also analyzed by PathWeather and compared against a gold standard de novo assembler and the assembler used most often previously for *var* genes (Jespersen et al. 2016), SPAdes (v3.11.0) (Bankevich et al. 2012). SPAdes can be run with a standard or "careful" mode, which maps sequences back to the assembled contigs to try to error correct them. Results from both the default mode and the careful mode along with PathWeaver are shown in **Table 4.1.** In samples with more than 1 copy of *var2csa*, the default SPAdes program

created inaccurate contigs in almost all but two of the samples. While SPAdes careful mode improves its results it still fails on the majority of the samples with copies of *var2csa* >=3, while PathWeather accurately reconstructed contigs in all samples with three or less copies and created only one false contig in the samples with 4 *var2csa* copies.

Field Samples

We then sought to better determine the global diversity of *var2csa* using available whole genome sequence data from global isolates. Approximately 2,900 field samples were processed yielding a total of 743 UpsE-ID5 sequences and an average of 1,800 sequences on sub-regions, see **Table 4.2** and **Table 4.3** for a breakdown of total sequences found per geographical region and *var2csa* sub-regions. Rarefaction curves were created for each defined region broken down for each geographical region, see **Figure 4.1.** The rarefaction curves only reached saturation in Southeast Asia suggesting a significant level of diversity in the African regions yet to be described.

Given the high number of variants described, we assessed for population structure using Principal Components Analysis (PCA). When evaluating the complete amino acid sequence of NTS-ID5 contigs (3D7 codons 1-2481), we identified two major population clusters (**Figure 4.2**). **Figure 4.3** shows that most of the structured amino acid variation (positions with the highest loading values for PC1 and PC2) is mostly contained in two regions within the MCBD (the ID1 hypervariable region and DBL2 hypervariable region). Of note, while these regions showed an excess of variation, there remained a high level of diversity along the entire region of the gene analyzed. We then evaluated how individual sub-domains of the protein impacted the population structure by generating PCAs based upon their amino acid sequence (**Figure 4.4** and **Figure 4.5** of all domains). The MCBD

(3D7 codons 373-999) (**Figure 4.4a**), combined hypervariable region of ID1 and DBL2 (3D7 codons 392-624) (**Figure 4.4b**), ID1 hypervariable region (3D7 codons 392-568) (**Figure 4.4c**) and DBL2 hypervariable region (3D7 codons 585-624) (**Figure 4.4d**) broke into 4, 6, 4 and 3 groups respectively and PCA of MCBD outside of the polymorphic region shows no structure (**Figure 4.6**). PCA of the regions outside of these shows little to no structure (**Figure 4.5**) with the exception of DBL1. Interestingly, all major groups were found across all geographic regions and were also found across years of collection consistent with a semi-stable population for these most polymorphic regions (**Figure 4.7**).

The number of sequences that make up each group and the amount of amino acid conservation is summarized in **Table 4.4**. Though the PCA plot shows fairly tight clustering in this region, amino acid conservation for within groups averages at 55% ranging from 40% to 62% though that is much higher than the 18% conservation when all sequences within this region are considered.

UpsE Open Reading Frame

Among var genes, var2csa is distinct in terms of its upstream sequence, UpsE. All 1559 sequences collected for the UspE open reading frame had the start codon and stop codon conserved and 79% of its 119 codons are perfectly conserved. The translated protein doesn't match any other proteins when protein blasted on the NCBI website. Though it has been found that this region is present in the VAR2CSA mRNA transcript (Lavstsen et al. 2003) it's not clear what function it serves or if it is also translated. This open reading frame is also conserved in the UpsE region of the *P. reichenowi* ortholog of *var2csa*.

Copy number variation in var2csa

To examine the extent of copy number variation, we limited our analysis to field samples representing monoclonal infections. Monoclonal infections were determined by running PathWeaver on 300 non-overlapping additional hypervariable 200 bp window in 230 single copy genes. A sample was then classified as monoclonal if data was recovered for 200 or more of these loci and if PathWeaver constructed only a single haplotype. This identified 1514 monoclonal samples. Since samples were observed to have non-uniform coverage which biased copy number conformation, we then eliminated samples using a t-SNE analysis to cluster samples with uniform coverage away from the non-uniform sample (Supplemental Figure **Figure 4.8**) and additionally eliminated samples with a coverage standard deviation of greater than 20 and mean coverage of less than 50. This left 525 high quality coverage monoclonal samples for copy number variation analysis. Of these samples, 373 (71%) had a singular contig constructed and the coverage for each matched mean base coverage in areas with similar GC content across the genome. Due to read sizes not always being long enough to span areas of conservation, full NTS-ID5 contigs were not always possible to generate, leading to several different contigs (see **Figure 4.9**). Mean base coverage was again normalized to the mean base coverage of genomic regions with similar GC content. For each sample with increased copy number, each copy var2csa had a unique haplotype. Copy calls were then summed by geographic regions and collection year and shown in **Figure 4.10**. Samples from South America were only found to have a single copy of var2csa across all years sampled there (2009-2012), while samples from South East Asia had samples with up to 3 copies and African samples had evidence of up to 5 copies. This distribution is similar to other studies of copy number variation in the *Plasmodium falciparum*

genome (Cheeseman et al. 2016). Also, another lab strain other HB3, UGT5.1, was also found to have two copies of *var2csa*.

The samples with 2 var2csa copies (n=107) were typed for their ID1 and DBL2 polymorphic hypervariable regions to determine if they contained different types, typing was done with what PCA group on the ID1 and DBL2 polymorphic regions PCA analyses **Figure 4.11**, **Figure 4.12**. These samples almost all have the at least 1 copy of ID1 type 2 with the majority of samples (62.2%) having a different ID1 type in the other copy of var2csa. Given the high prevalence of type 2 ID1 hypervariable regions in the single var2csa copy monoclonal infections (87% of 383 infections), this type of ID1 region appears to be underrepresented in parasites with more than one var2csa copies (66% of copies) (Chi-squared test X-squared = 53.7, df = 3, p-value = 2.64e-11). There did not appear to be a significant different in the number of DBL2 types (X-squared = 6.68, df = 2, p-value = 0.0708)

We used Pf3k's PacBio assembled genomes to then confirm the locations of multiple copies of var2csa on these genomes. The genomic location of the HB3 duplicate was confirmed on chromosome 1 when extracting from the Pf3k assembled genomes. Two of the clinical isolates were also found to have multiple VAR2CSA genes. One isolated, PfSN01 from Senegal, had two copies on chromosome 12 and the other isolate, PfTG01 from Togo, had 2 copies on chromosome 12 and 2 on chromosome 8 in close proximity, all of which had an intact UpsE region, see **Figure 4.13**. The relatedness of the four copies in PfTG01 ranged from 90.5% to 93.8% and the relatedness of the two copies in PfSN01 was 92.0%. Previous studies have suggested that the possible location of additional VAR2CSA copies could be on chromosome 8 which would consist with what is seen here (Sander et al. 2011).

Discussion

Var2csa, which mediates the binding of malaria parasites to the placenta, is an important potential vaccine candidate to prevent malaria in pregnancy. Antibodies to var2csa have been found to be protective against malaria-associated poor birth outcomes(Rogerson et al. 2007; Ataíde, Mayor, and Rogerson 2014). The study of *var2csa* is obstructed by its high diversity which prevents it from being studied with traditional reference based variant calling as it does not map well to a single reference. Another hurdle to studying *var2csa* using short read sequencing libraries is the fact that it has multiple copies and that many *P. falciparum* infections are polyclonal which means special care has to be taken when attempting to apply genome assembly approaches that unique copies are not improperly stitched together to form false sequence.

Here we introduced a novel algorithm, PathWeaver, for extracting local haplotype sequence from even highly diverse regions of a genome and within polyclonal infections. The need for this algorithm arose from current assembly tools not being specifically designed for this purpose. While SPAdes does well on monocopy samples, which is what it was designed for, it does create false haplotypes even on careful mode when dealing with polyclonal samples. Other assembly type programs were tested as well, Trinity, Velvet, SSAKE, megahit among others and all suffered when it came to polyclonal samples (data not shown) for various reasons ranging from over collapsing variation to creating false haplotypes to not allowing output contigs to both contain regions of conserved sequence. While some of these problems had the potential to be circumvented with writing programs to wrap the assemblers in various ways a greater benefit was seen in creating a custom assembler specifically designed to handle cases of polyconality and that could be more

easily adapted for various scenarios it was needed for. While there do exists programs that were created to handle local haplotype assembly of polyconal infections (ShoRaH (Zagordi et al. 2011), ViQuaS (Jayasundara et al. 2015), etc) these all work directly with only the mapped portions of sequences rather than the full query sequence which, as shown above with 30% of *var2csa* sequences being clipped off, wouldn't work for regions like *var2csa* and the concept of PathWeather was conceived with special interest in regions like var2csa, other var genes, and other *Plasmodium* regions that contain DBL domains or other binding domains that often show polymorphic allelic types(Crosnier et al. 2016; Ware et al. 1993; McColl and Anders 1997; Pearce et al. 2004).

By utilizing *in silico* simulation of 30 different *var2csa* variants we have proved that our iterative recruitment method of using 3D7 *var2csa* and subsequent recruitment of unmapped reads is adequate to gather enough UpsE-ID5 *var2csa* reads to be assembled into the expected sequence. Followed by tests on lab strains which had been whole genome shotgun sequence showed that this recruitment method only assembled the expected *var2csa* and does not recruit other *var* sequences. What aids in this endeavor to recruit only *var2csa* sequence is the utilization of mapping all reads to the 3D7 genome and pulling only the sequences that map to the 3D7 *var2csa* which due to the nature of mapping reads via local alignment and soft clipping if a portion of the read matches a region then it will be recruited unless it matches another region better. Though *var2csa* is fairly unique among the *var* genes it still shares some homology blocks with other *vars* as well as other DBL proteins like EBA-175 and so if only *var2csa* was used to recruit reads it could improperly recruit reads from other regions while using the whole 3D7 genome will help recruit these similar reads to their proper regions because they will more closely match those regions. However, it was observed that from the *in silico* simulations that approximately 1-3% of sequences

were recruited to other 3D7 regions including *var* genes and an algorithmic improvement could be to incorporate checking other regions once initial contigs have been assembled but based on our results here it was shown that this wasn't necessary to assemble the expected sequence. It might also be tempting to utilize multiple *var2csa* sequences in the initial recruitment but this increases the potential danger of recruiting sequences other than var2csa and again we with our results here we have shown that such an approach is not necessary. In addition, by mapping to the 3D7 reference genome for the initial recruitment also allows the study of multiple regions from the same alignment file, which can be quite large and could potentially hamper the investigation of several regions at once if multiple alignments had to be created for each region. PathWeaver has allowed us to extract a large number of sequences for *var2csa*.

PCAs of the entire gene show little structure (**Figure 4.2**). However, a higher degree of structuring is observed when focusing on the MCBD and its polymorphic regions which demonstrates 4 major groups and 2 minor groups (**Figure 4.4**). Thus, much of the structured diversity in var2csa is found within the MCBD. This is consistent with evidence that the MCDB is the principal CSA ligand and that antibodies to the MCBD are particularly protective (Rogerson et al. 2007; Ataíde, Mayor, and Rogerson 2014). The MCBD regions also show the least amount of conserved amino acids and the highest mean expected heterozygosity for the gene (**Table 4.3**).

Balancing selection is a phenomenon that selects for diversity, especially in immune epitopes; the more diverse an epitope, the more likely the parasite is able to survive and reinfect a host with a previous infection especially if cross-strain reactive antibodies are not able to be formed. This is especially true for infectious agents that don't induce lasting immunity and leads to individuals being infected multiple times which leads to a strain's

frequency being inversely correlated with its survivability (Lipsitch and O'Hagan 2007). There is substantial evidence for for balancing selection in malaria antigens especially for blood stage antigens (Weedall and Conway 2010). All 4 major groups and to a certain the 2 minor groups are observed at a similar frequency across time and space (**Figure 4.10**) suggests that the same balancing selection forces are occurring independently on different continents.

Gene duplication in var2csa has been previously reported (Sander et al. 2009, 2011). Here, using the PathWeaver algorithm we were able to detect multiple copies of *var2csa* in monoclonal field samples found that each copy had approximately mean base coverage. When multiple copies do exist in a single genome, they were always found to be unique from each other which is consistent with previous findings (Sander et al. 2009). We were able to utilize chromosome level assemblies provided by MalariaGen's Pf3k project (https://www.malariagen.net/projects/pf3k) to confirm that *var2csa* has a conserved chromosome 12 loci and to show that the genomes of two field isolates had two tandem copies on chromosome 12 for one isolate and the other isolate had two tandem copies on 12 and 8 which is also consistent with previous finding suspecting the possible location of additional var2csa copies being on 8 (Sander et al. 2009). As more chromosome level genome assemblies become available the possible locations of these copies can further characterized. It's been postulated before that the multiple copies could explain the polymorphic types seen within *var2csa* (Sander et al. 2009), however we have observed here that all types for both ID1 and DBL2 appear in the monocopy samples so though the multiple copies are likely helping drive this diversity it doesn't appear that the conserved chromosome 12 loci is associated with just one type.

We have only scratched the surface for analyses that could be done here. However, conventional tools and metric are not easily applied to a gene like *var2csa* with its complex evolutionary history, high rate of recombination, high diversity preventing the ability to use one sequence as a reference and its multiple copies. With having extremely divergent types there is no single good reference for sequences to be compared to which is the basis/requirement for many traditional measures of diversity and other population structure analyses. Also, a 3D structure could greatly inform the information gained from the sequence variation gathered here to see if variation is buried or forms pockets. There is currently no 3D structure available for *var2csa* but the amount of sequence gathered here could aid in the simulation of one.

Methods

PathWeaver

PathWeaver represents a novel iterative multistep assembly method that allows accurate local assemblies of highly variable genes. Raw sequences are aligned to a reference genome using BWA-MEM (http://bio-bwa.sourceforge.net/) with default parameters. Extracted reads are then processed through a custom graph based method described below. Extracted reads are all oriented to the either the plus or negative strand, depending on input settings, so that final contigs are all oriented in the same direction. Recruited reads are then k-mer indexed at a certain k-mer size, default 40, to create nodes, nodes that fall below an occurrence cut off, default 5, are remove. Edges are then added to connect the nodes using a method called “threading” (J. R. Miller, Koren, and Sutton 2010) where nodes are connected if the k-mers occur adjacent in the input reads as opposed to a

classical approach of simply connecting k-mer nodes when their suffix and prefix match perfectly. Once de novo contigs have been constructed, the initially unmapped sequences are then aligned against the de novo assembled contigs using BWA-MEM (http://bio-bwa.sourceforge.net/) again. Graph assembly is repeated using the reads initially pulled down and these newly recruited reads to create new contigs. The still unmapped sequences are then aligned again to the new contigs to recruit more of the unmapped sequences. This is done iteratively until there are no newly recruited sequences or a max iteration number is hit (default of 20).

Once the final iteration is done the final sequences are then trimmed to the region of interest. The number of final contigs and whether they span the whole region will be dependent on the size of the region, the depth, the amount of variation present if there are more than one unique copy of the region of interest, and the size of the read length of the input data. For example, if unique copies share a region of conserved sequence longer than read length than the variation flanking the conserved region cannot be stitched together or if a portion of the region of the interest fails to get sequenced than the output will be several contigs that covered the region. Another scenario where a full length contig might be possible is if there is a tandem repeat in the region which is longer than the read length and therefore the size of the repeat cannot be easily determined, several contigs will be reported even when there is only one unique copy. See **Figure 4.14** for a visual representation of the read recruitment strategy.

var2csa Assembly

The untranslated upstream region UpsE to the inter domain 5 (ID5) region of var2csa (UspE-ID5, genomic position Pf3D7_12_v3 49360-57446 in 3D7 (v3)

(http://plasmodb.org/common/downloads/release-34/Pfalciparum3D7/fasta/data/PlasmoDB-34_Pfalciparum3D7_Genome.fasta)) was chosen for the *var2csa* assembly to ensure the highest recovery of data. Exon 2 of *var2csa*, like most *var* genes, shows a high degree of similarity to other *var* exon 2 sequences and has a high potential to recruit other *var* sequences and so was avoided. DBL6 is split between exon 1 and exon 2 and was also avoided recruiting sequences that extend into the intron, which has hard to assembly sequences with long tandem repeats. For samples with multiple copies of *var2csa*, full length UspE-ID5 reconstruction is not possible due to long conserved regions that reads are unable to span and the unique path cannot be determined. Partial sequences were used for subregion analyses if they covered the entire length of the analyzed region.

In silico Simulations of *var2csa* UpsE-ID5 sequences

In order to test the PathWeather algorithm, we simulated shotgun sequencing of UpsE-ID5 *var2csa* sequences. We collected UpsE-ID5 sequences by using MalariaGEN's 15 Pacbio chromosome level genome assemblies for 5 lab strains (GB4, 7G8, DD2, HB3 and IT/FCR3) and for 10 clinical isolates (https://www.malariagen.net/projects/pf3k). UpsE-ID5 of 3D7 *var2csa* was extracted from these genomes by determining *var2csa's* location using LASTZ (Harris 2007). Additionally, the *var2csa* sequences from a previous study on *var* genes which collected *var* sequence from NCBI's BLAST and available lab genome assemblies (Rask et al. 2010). A total of 30 unique *var2csa* sequences were collected and each were used to simulate a shotgun 2x100 Illumina sequencing run with approximately 40 reads per base coverage.

Parasite Whole Genome Shotgun Sequencing Data

Data was collected from several publicly available studies (Baniecki et al. 2015; Cerqueira et al. 2017; Parobek et al. 2017; Dara, Drábek, et al. 2017; Kumar et al. 2016) by parsing the SRA database and by using MalariaGEN's Pf3K (https://www.malariagen.net/projects/pf3k) and Broad's 100 genome project (Plasmodium 100 Genomes initiative, Broad Institute (https://www.broadinstitute.org)). The number of samples collected for each country was Bangladesh=50, Cambodia=663, DRC=113, French Guiana=58, Ghana=605, Guinea=100, Laos=85, Malawi=269, Mali=119, Myanmar=60, Nigeria=5, Senegal=137, Thailand=536, The Gambia=65, Uganda=11, Vietnam=97. In addition to the field samples, datasets contained the following lab strains, DD2, GB4, W2, IT/FCR3, 3D7, Tanzania (2000708), UGT5.1, 7G8, FCH/4, CAMP/Malaysia, MaliPS096_E11, NF135/5.C10, NF54, Santa Lucia, Palo Alto/Uganda and Vietnam Oak-Knoll (FVO). See below for a description of each study.

Pf3k

MalariaGEN's Pf3K sequence reads (data release 5) of 2,512 whole genome sequencing samples were downloaded using their SRA accession numbers (https://www.malariagen.net/projects/pf3k). Data represents samples collected from 14 different countries across Africa and Southeast Asia {Supplemental Table}. The majority of samples were 2x100 paired end Illumina sequencing. The Pf3k data also consists of 28 lab control various mixtures of 3D7, DD2, 7G8, and HB3 with the strains mixed with frequencies range from 1-99% (6 mixtures of 3D7/DD2, 3 mixtures of DD2/HB3/7G8, 16 mixtures of

HB3/7G8, 1 monoclonal of HB3, and 1 monoclonal of 7G8). Pf3K contains multiple instances of some samples and these were removed.

Other Lab Strains

Several more monoclonal lab strain sequencing samples were also found by querying the SRA database. The strains were DD2 (ERR663287), GB4 (ERR027100), IT/FCR3 (ERR713965), and W2 (ERR663245) and 10 3D7 samples (ERR043381, ERR043382, ERR044266, ERR047177, ERR047178, ERR047179, ERR047184, ERR047185, ERR047186 and ERR047187).

Broad 100 Genomes Project

The data from The Broad Institute's 100 Genome project for Plasmodium (Plasmodium 100 Genomes initiative, Broad Institute (broadinstitute.org)) has produced whole genome sequencing for the following lab strains Tanzania (2000708), UGT5.1, 7G8, FCH/4, CAMP/Malaysia, MaliPS096_E11, NF135/5.C10, NF54, Santa Lucia, Palo Alto/Uganda and Vietnam Oak-Knoll (FVO). The majority of lab strains were sequenced in triplicate with two libraries having a target insert size of 180 and 1 library with a target insert size of 5000. Whole genome sequencing was also produced for 44 samples from three different countries, 22 from French Guiana, 11 from Mali and 11 from Uganda. The French Guiana samples were sequences similarly to the lab strains described above while the Mali and Uganda samples were sequenced only once with a target insert size of 180. All libraries were sequenced by Illumina 2x100 paired ends.

Baniecki et al. 2015

An additional 34 samples from a longitudinal study in French Guiana from South America (Baniecki et al. 2015).

Cerquerira et al 2017

An additional 179 samples from a longitudinal study in Thailand were used though the majority of these samples were captured used hybrid capture which avoid the var gene regions so some samples lacked *var2csa* sequences (Cerqueira et al. 2017).

Parobek et al 2017

Data from a study on the effect of artemisinin partner drug usage had 93 clinical samples from 3 different regions in Cambodia, see paper for further details (Parobek et al. 2017).

Dara et al 2017

Data from a study on constructing var genes by utilizing Pacbio assemblies combined with Illumina paired end sequence had 12 samples from a village in Mali, see paper for further details (Dara, Travassos, et al. 2017).

Kumar et al 2016

Five samples from a study looking at Plasmodium falciparum diversity in India were also used (Kumar et al. 2016).

Determining Monoclonal Samples

In order to investigate copy number we needed to determine which samples were monoclonal. We identified 300 hypervariable non-overlapping 200bp windows within 230 single copy genes spread across all the chromosome of the Plasmodium genome. These were identified by interrogating the genomes of known laboratory strains to find small hypervariable regions that uniquely mapped back to the reference strain. We then determined that these windows had reliable coverage across the clinical datasets, then using PathWeather to conduct local reconstruction. A region was only considered if all contigs constructed spanned the whole region and not just a sub-portion of the region as these could represent failed constructions due to having more than one clone. Contigs also had to consist of at least 98% of the recruited reads to a region to ensure that all possible variants were being assembled. A sample was then classified as monoclonal if data was recovered for 200 or more of these 230 genes (some genes have more than 1 of the 300 windows) and if PathWeaver constructed only a single haplotype. These windows had an expected heterozygosities ranging from .53 to .98 (mean 0.68).

Analysis Programs Used

Other analysis methods: Rarefaction curves generated by R package vegan (v2.4-6)(Oksanen et al. 2018); PCAs on multiple protein alignments were generated by custom c++ scripts using the method described in (Wang and Kennedy 2014) and using R's (R version 3.4.3) prcomp followed by group clustering by Hierarchical DBSCAN (Ricardo J G, Moulavi, and Sander 2013).

Tables

Table 4.1: Pf3k Control Assembly Programs Results

Program	# of var2csa copies	Error Free Samples	Total Samples	Average # of False Contigs
PathWeaver	5	5	5	0
PathWeaver	2	7	7	0
PathWeaver	3	16	16	0
PathWeaver	4	2	3	0.67
Spades	5	5	5	0
Spades	2	1	7	1.29
Spades	3	1	16	5.81
Spades	4	0	3	9
Spades-careful	5	5	5	0
Spades-careful	2	6	7	0.29
Spades-careful	3	2	16	2.75
Spades-careful	4	0	3	6.33

Table 4.2: Reconstructed Sequences Count Per Region

Region*	UpsE-ID5	NTS-ID5 (Codon1-Codon2481)**	ID1-DBL2x-ID2a (MCBD) (Codon372-Codon999)**	Polymoprhic Region in MCBD (Codon391-Codon624)**
South America	44 (7)	52 (8)	54 (7)	55 (7)
West Africa	158 (142)	246 (205)	331 (240)	856 (506)
Central Africa	23 (23)	34 (32)	40 (38)	111 (100)
East Africa	54 (52)	68 (63)	82 (72)	274 (210)
India	4 (2)	4 (2)	5 (2)	6 (2)
South East Asia	456 (106)	518 (115)	644 (132)	1079 (172)
Total	739 (332)	922 (420)	1156 (472)	2381 (906)

Number of total sequences collected with number of unique sequences collected in parentheses

* Countries per Region: Central Africa = Democratic Republic of the Congo; East Africa = Malawi,Mozambique,Uganda; India = India; South America = French Guiana; South East Asia = Bangladesh,Cambodia,Laos,Myanmar,Thailand,Vietnam; West Africa = Ghana,Guinea,Mali,Nigeria,Senegal,Gambia

** Start amino acid (AA) codon to the beginning of DBL6

*** Minimal binding domain spanning ID1-DBL2x-ID2a

**** Polymorphic Region within minimal binding domain, includes regions from ID1 to DBL2

Table 4.3: Counters Per Domain

region	3D7 Codon Start	3D7 Codon Stop	length	# of Sequences	# of Unique Sequences	# of Singlets	Max Unique Count***	# of Samples****	# of Positions >=95% Conserved *****	% of Positions >=95% Conserved *****	Mean He(SD)
Upse-ID5*	49361	57446	8086	743	332	261	60	730	6076	75.14%	0.241 (0.202)
UpsE ORF**	1	119	119	1553	197	81	223	1462	95	79.83%	0.213 (0.177)
NTS-ID5	1	2481	2481	904	405	309	63	904	1438	57.96%	0.317 (0.221)
NTS	1	65	65	2864	693	353	139	2109	40	61.54%	0.296 (0.225)
DBL1	66	356	291	1694	612	397	73	1480	149	51.20%	0.301 (0.227)
ID1	357	568	212	2558	936	591	110	1936	31	14.62%	0.365 (0.196)
Minimal CSA Binding Domain	372	999	628	1145	465	331	72	1095	279	44.43%	0.328 (0.224)
Minimal CSA Binding Domain Polymorphic Region	391	624	234	2366	885	569	108	1869	44	18.80%	0.37(0.197)
ID1-Polymorphic	391	572	182	2648	914	554	112	1992	20	10.99%	0.367 (0.191)
DBL2	569	916	348	1286	509	359	77	1196	213	61.21%	0.252 (0.245)
DBL2-Polymorphic	579	624	46	3313	306	107	210	2272	17	36.96%	0.419 (0.208)
ID2	917	1331	415	1143	440	313	72	1118	229	55.18%	0.34(0.206)
DBL3	1332	1646	315	1096	414	298	75	1095	230	73.02%	0.29(0.202)
ID3	1647	1715	69	1751	166	72	165	1594	54	78.26%	0.337 (0.222)

DBL4	1716	2001	286	1135	415	284	76	1130	209	73.08%	0.156 (0.195)
ID4	2002	2150	149	1413	455	283	67	1294	115	77.18%	0.199 (0.204)
DBL5	2151	2430	280	1564	573	384	81	1412	144	51.43%	0.256 (0.229)
ID5	2431	2481	51	3022	397	141	100	2158	23	45.10%	0.308 (0.259)

* 3D7 genomic location on chromosome 12 are given for the UpsE to ID5 region as it can't be fully translated

** Codon positions are for the open reading frame and not the VAR2CSA codons

*** The max number of times the same sequence was found

**** The number of samples sequences were recovered from, some samples contribute more than one sequence for a region if it has more than 1 contig spanning the region

***** 95% conserved meaning that for a given position 95% of the total sequences had the same amino acid/base

Table 4.4: Counts for MCBD PCA Groups

group	# of Sequences	# of Unique Sequences	# of Singlets	Median Unique Count	Max Unique Count	# of Samples	# of Positions >=95% Conserved	% of Positions >=95% Conserved	He
1	947	325	212	3	108	872	127	54.27%	0.975
2	52	27	15	2.5	6	50	95	40.60%	0.943
3	926	348	226	3	104	862	127	54.27%	0.979
4	96	11	5	10	38	96	147	62.82%	0.736
5	128	71	42	3	6	124	138	58.97%	0.98
6	230	105	65	3	18	227	145	61.97%	0.979

Figures

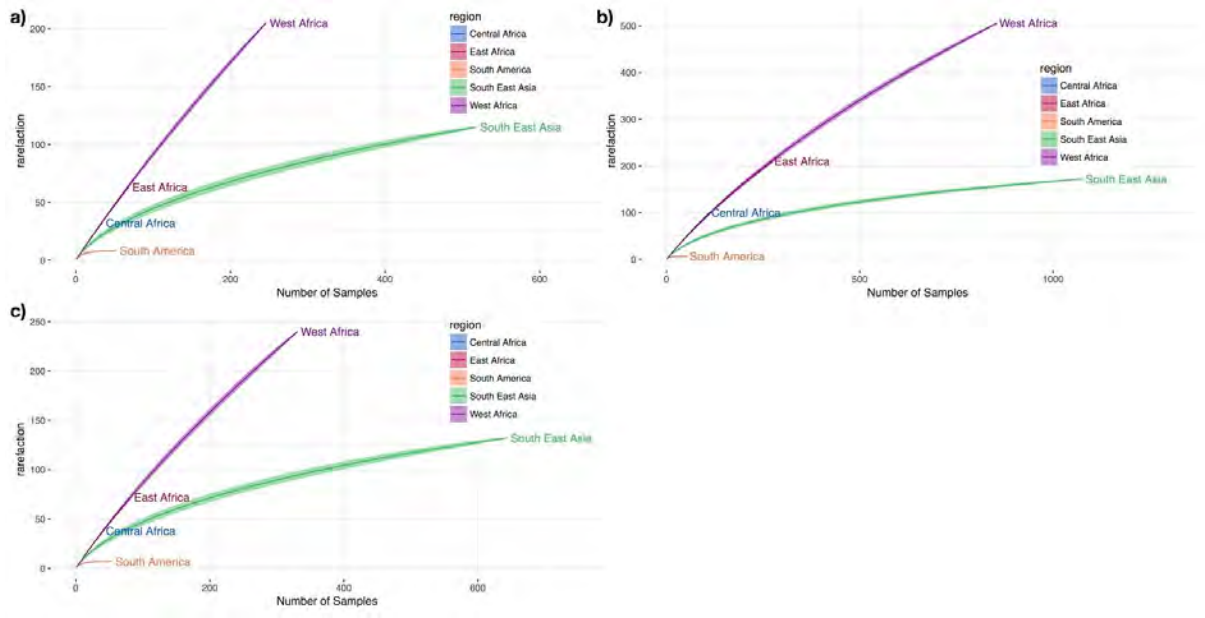


Figure 4.1: Rarefaction Curves

Rarefaction curves for the **a)** NTS-ID5 region, **b)** MCBD, and c) the polymorphic region in the MCBD. South East Asia shows signs of leveling out but the African regions are still very steep suggesting that there is still a high degree of diversity in Arica not yet documented.

Figure 4.2: PCA of NTS-ID5

PCA of the protein multiple alignment of the NTS-ID5 region. The nodes are colored by the groups based on the PCA of the polymorphic region in the MCBD. Lab strains are labeled. PC1's highest loading values fall mostly into the ID polymorphic region (**Figure 4.3a**)

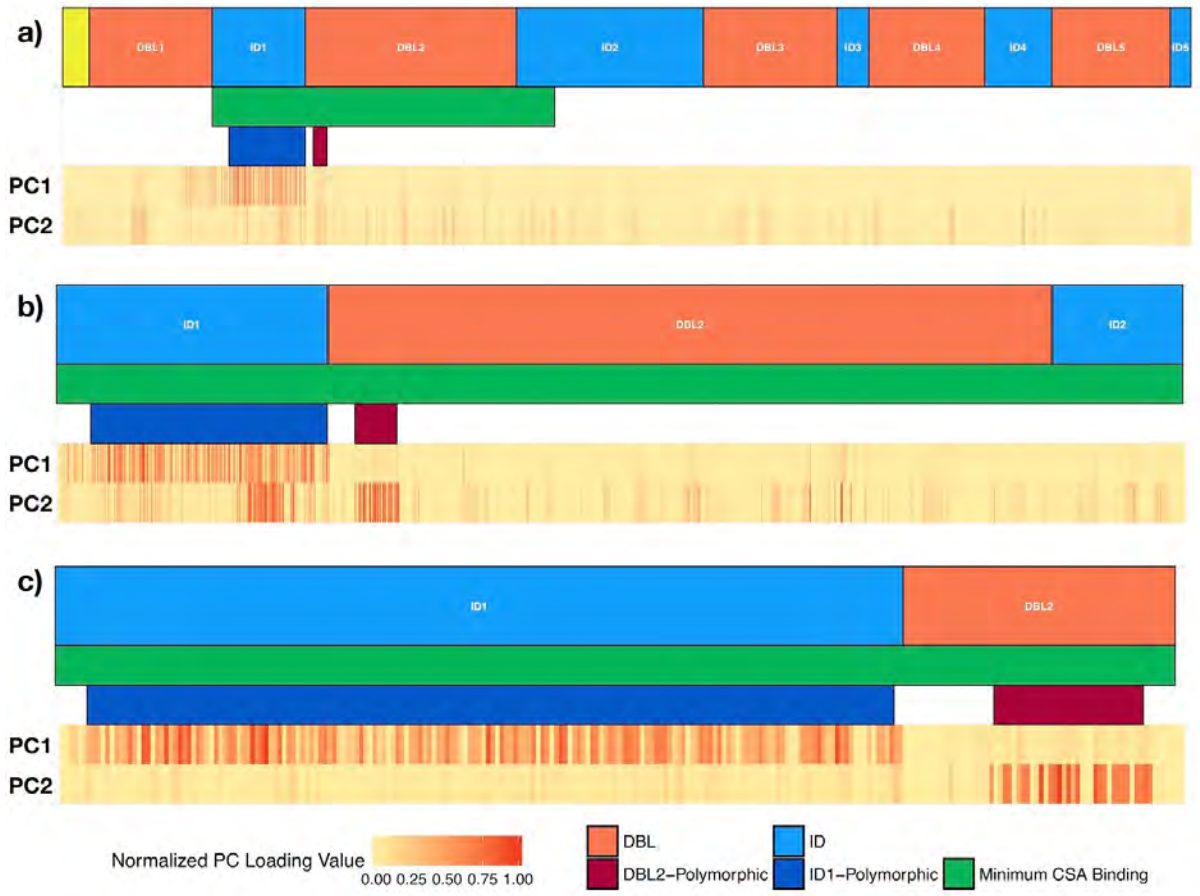


Figure 4.3: PC1 and PC2 Loading Values for NTS-ID5, MCBD, MCBD-Polymorphic

The PC1 and PC2 for the protein PCAs loading values normalized to the the max loading values for each, PC1 loading values are always the bar on top and PC2 loading values are always the bar on the bottom. **a)** shows the loading values for NTS-ID5, the highest values are PC1 fall within the ID1-Polymorphic **b)** shows loading values for the MCBD,the highest values for both PC1 and PC2 can be seen within the MCBD-Polymorphic region, and **c)** shows the loading values for the MCBD-Polymorphic region, showing that PC2 is being driven by the DBL2 polymorphic and PC1 is being driven by the ID1 polymorphic.

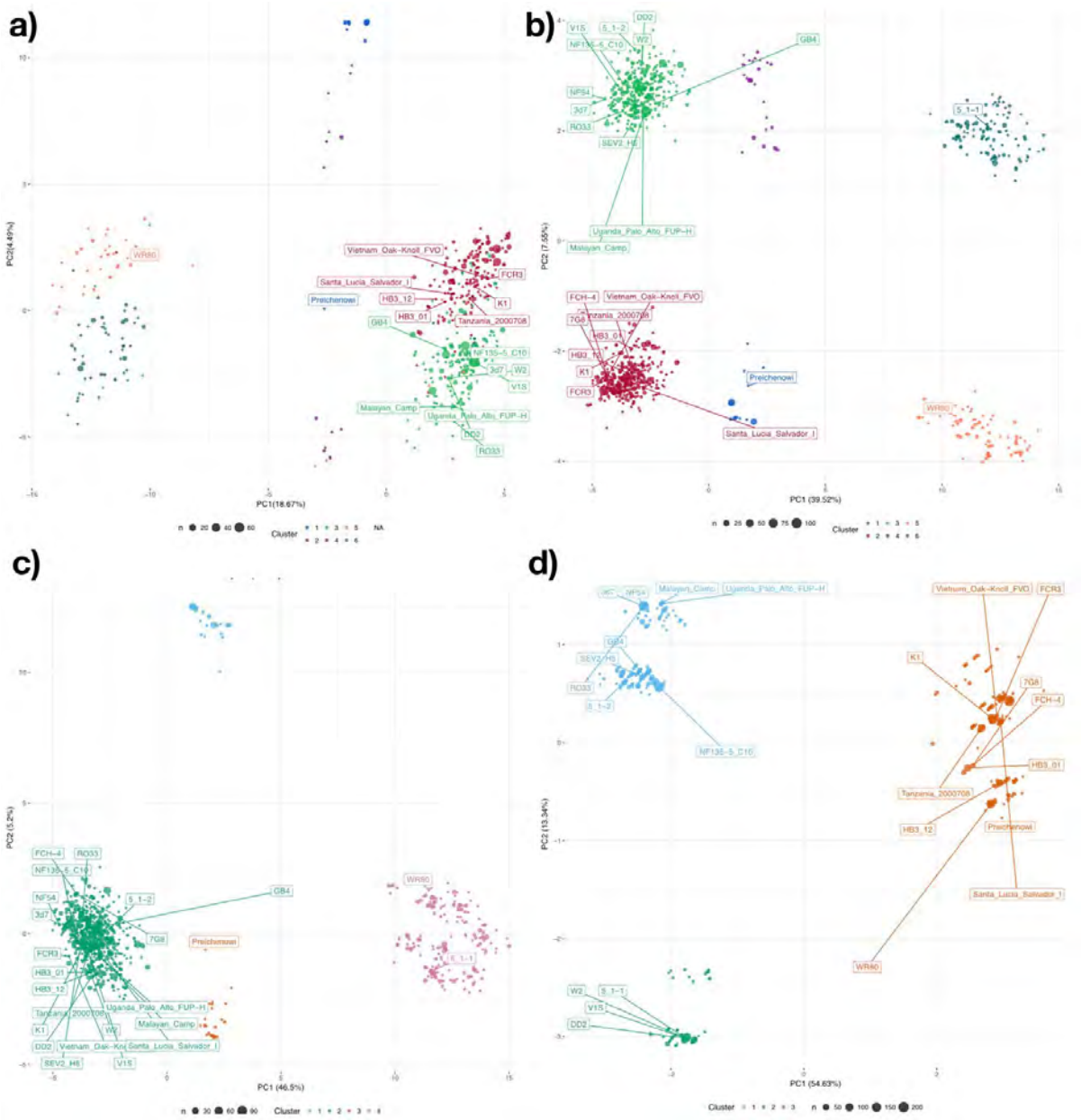


Figure 4.4: PCAs on the MCBD Domains

PCAs on protein multiple alignments for several domains within the MCBD. **a)** The full minimum CSA binding domain itself, it is colored by the clusters determined for the polymorphic region in **b)**, **b)** The polymorphic region of the MCBD made up of the ID1 polymorphic **c)** and the DBL2 polymorphic **d),** shows 4 major groups and 2 minor groups, **c)** the ID1 polymorphic region colored by HDBSCAN on its own region,splits into 4 groups **d)** the DBL2 polymorphic region, breaks up into 3 groups

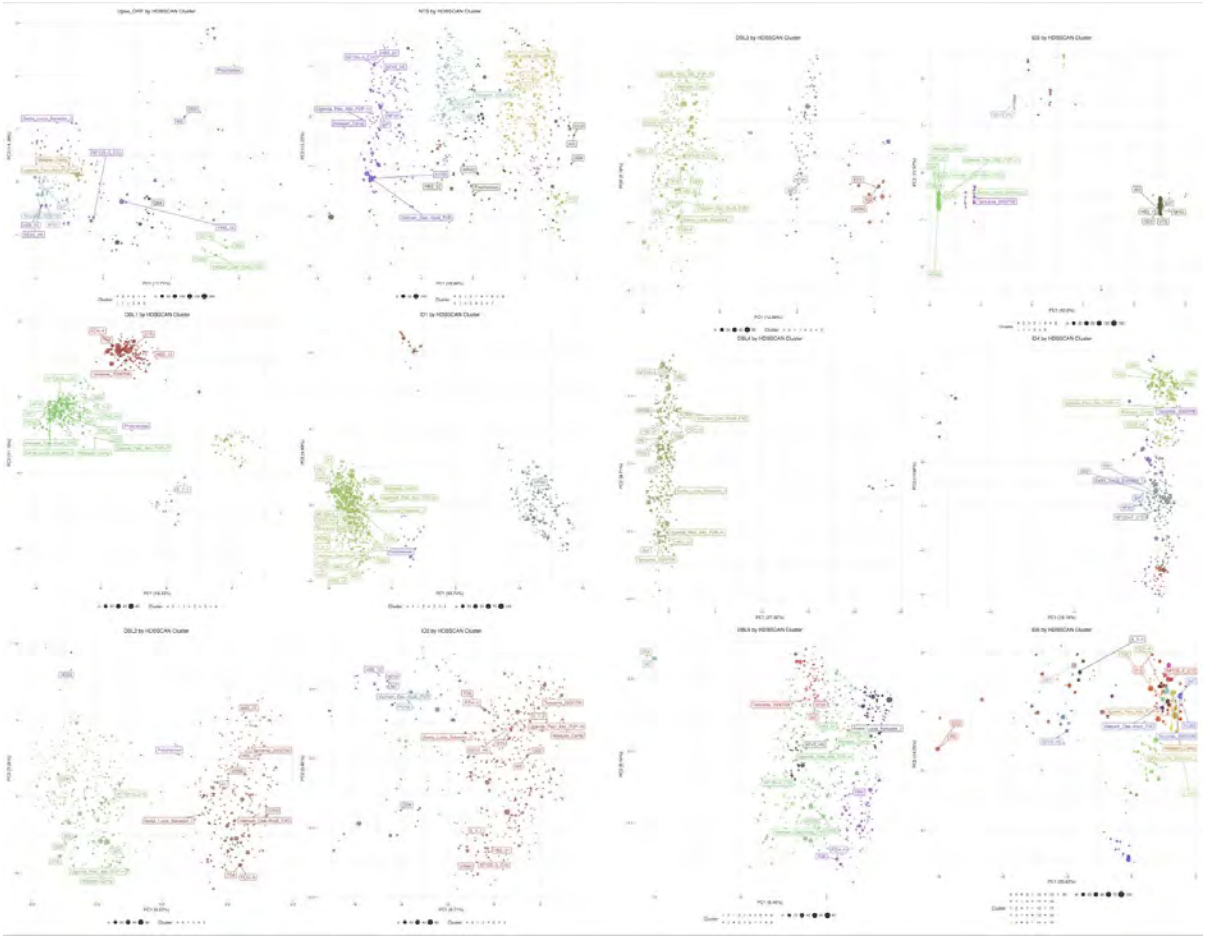


Figure 4.5: All Domains PCAs

PCA plots for each DBL and ID domains and the NTS and the UpsE ORF. Coloring is done by HDBSCAN on the PCA.

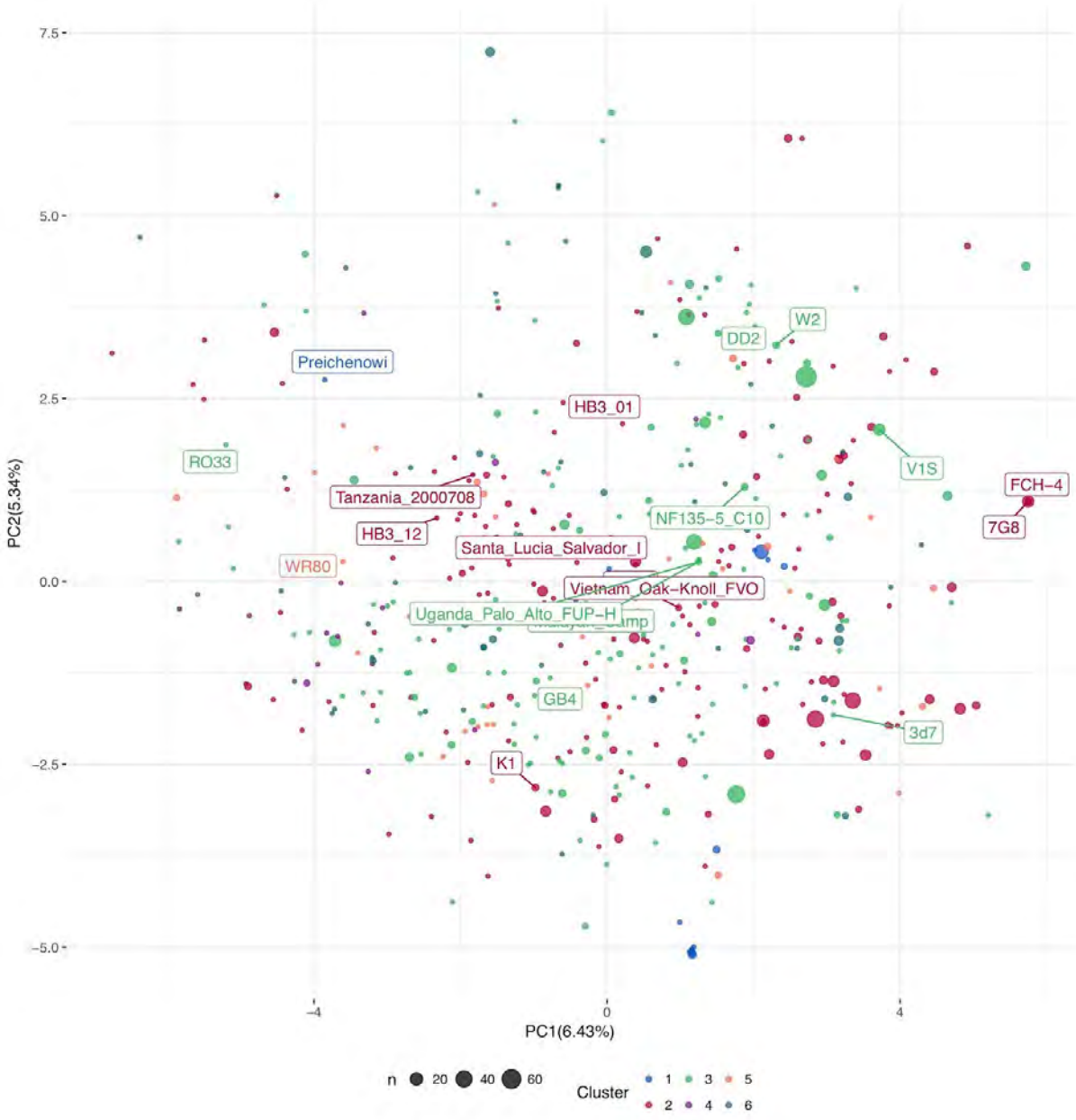


Figure 4.6: PCA of the Region Beyond the DBL2 within the MCBD

PCA of the region beyond the polymorphic region in the MCBD, cover most of the DBL2 region and the ID2 region within MCBD. No structure is evident from the plot showing that the majority of the structure in the MCBD is in the polymorphic regions in ID1 and DBL2.

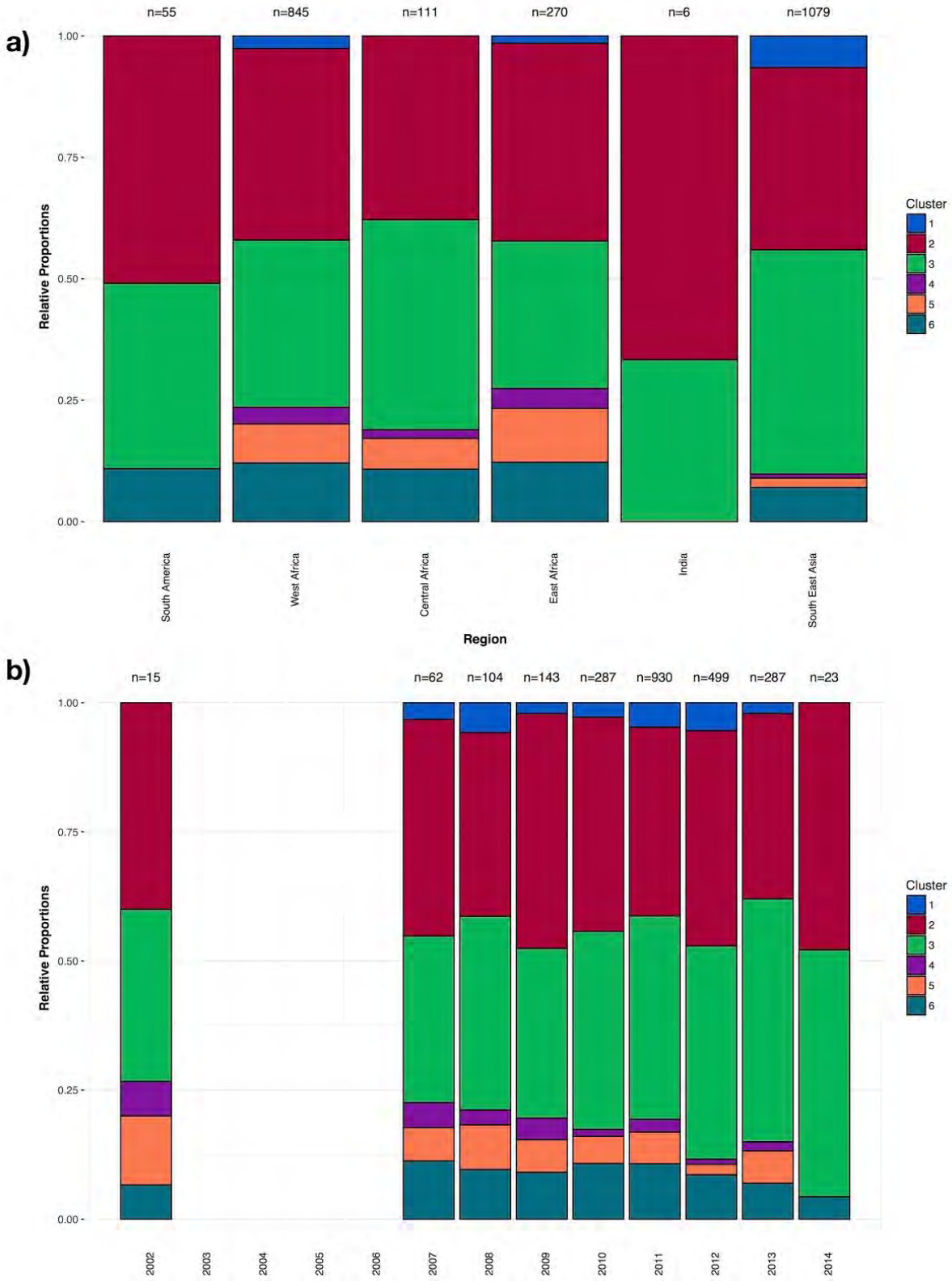

a)
n=55
n=845
n=111
n=270
n=6
n=1079
1.00
0.75
0.50
0.25
0.00
Relative Proportions
Cluster
1
2
3
4
5
6
South America
West Africa
Central Africa
East Africa
India
South East Asia
Region
b)
n=15
n=62
n=104
n=143
n=287
n=930
n=499
n=287
n=23
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014

Figure 4.7: MCBD Polymorphic PCA Group Counts

Counts across years and regions of the groups determined by the PCA on the polymorphic region in MCBD. The groups appear to stay stable across **a)** regions and **b)** years, suggestive of balancing selection keeping diversity.

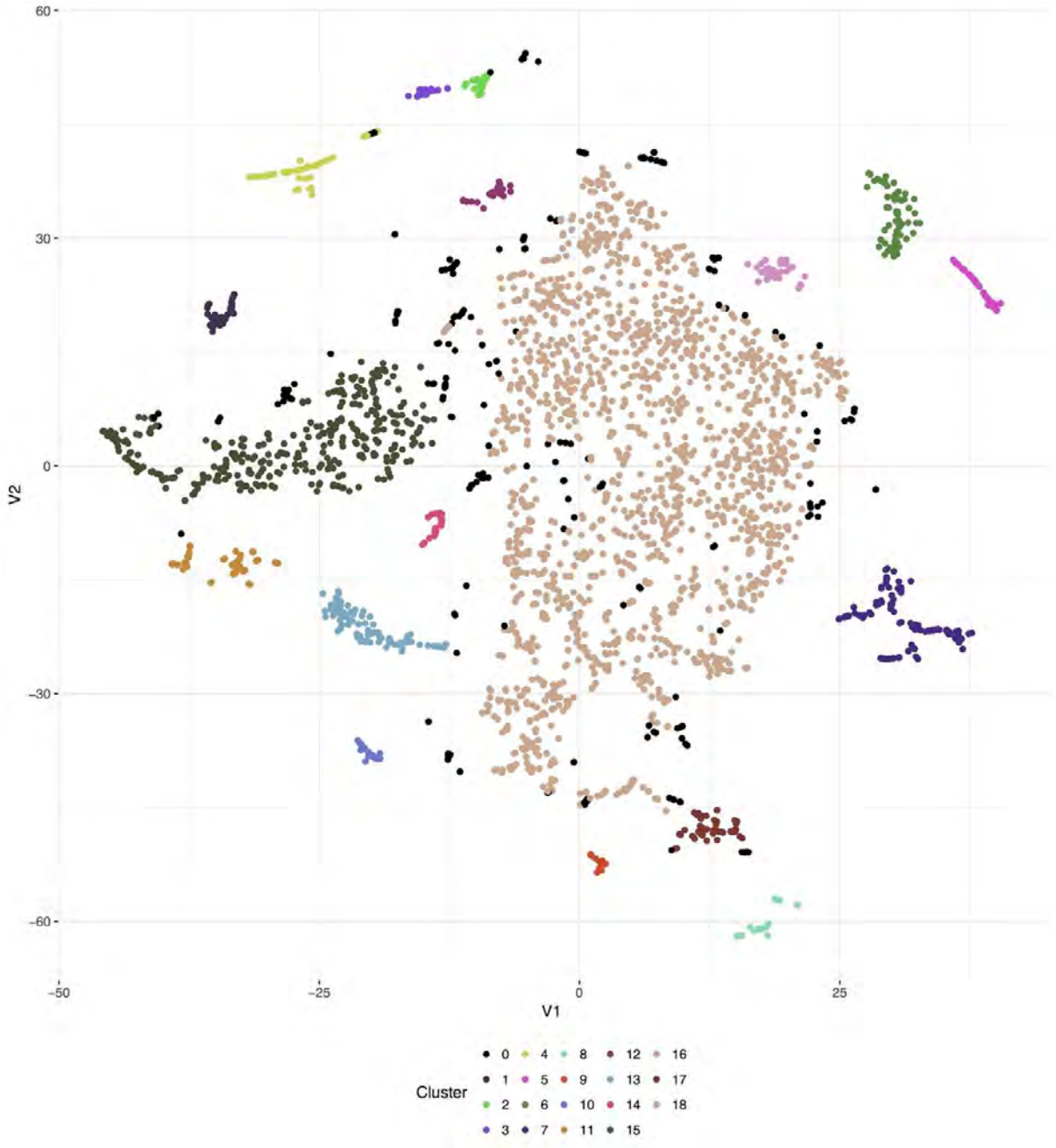


Figure 4.8: t-SNE of the Chromosome 12 Coverage

A TSNE off of the mean genome coverage of chromosome 12 of all monoclonal samples to select groups that were observed to have normal coverage.

Figure 4.9: Example of Assembly Output of 2 var2csa Copies Samples

Example of 4 monoclonal samples with evidence for two copies of *var2csa*. The relative positions of the domains are shown on the bottom. The length of the blocks indicate length and the height of the blocks indicate read depth. Blue rectangles have 1 tail and pink rectangles have 2 tails.

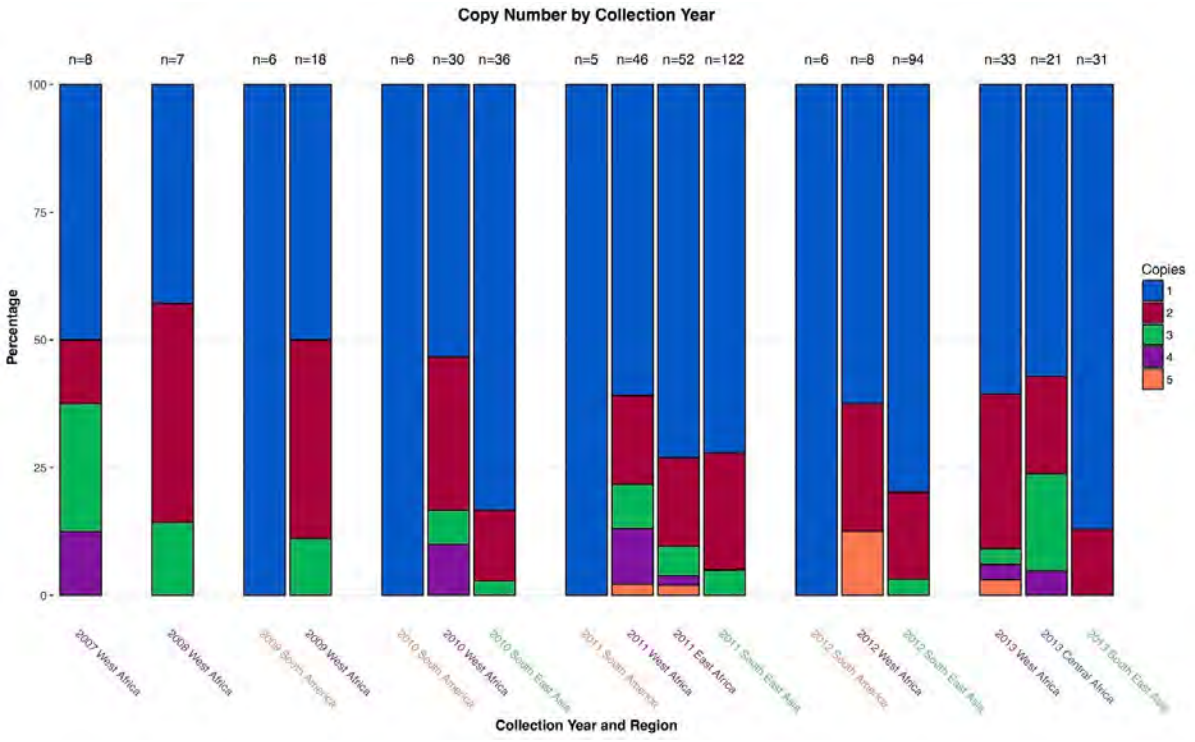


Figure 4.10: *var2csa* Copies Calls Across Time and Region

The x axis is both year and region, y axis is the relative amount of the monoclonal samples for each copy count. Bars are colored by the number of copies. The total amount for a given year and region is on the top of the bars.

Figure 4.11: ID1 Types Monocopy var2csa Samples

The y axis is the rounded var2csa coverage for this region divided by the mean base coverage and x axis is samples. The bars are colored by ID1 type determined in **Figure 4.4a.**

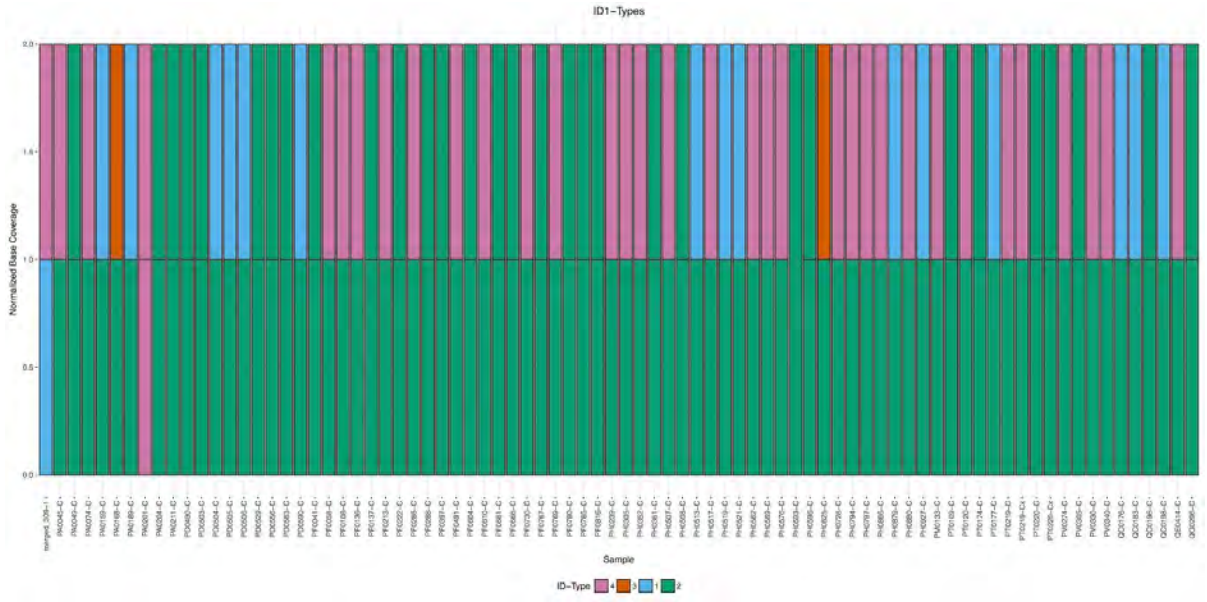


Figure 4.12: ID1 Types Two Copies var2csa Samples

The y axis is the rounded var2csa coverage for this region divided by the mean base coverage and x axis is samples. The bars are colored by ID1 type determined in **Figure 4.4a.**

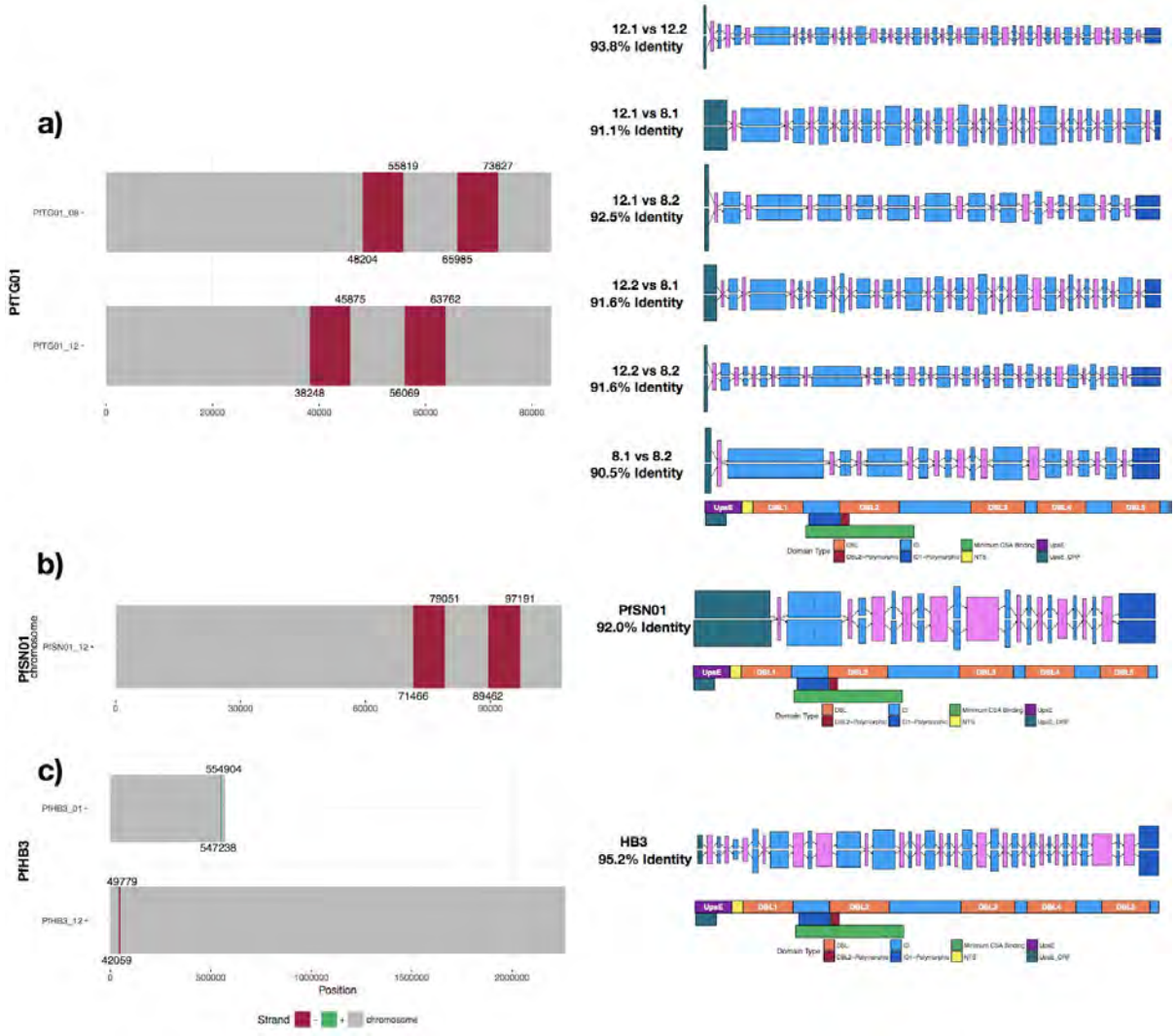


Figure 4.13: Locations of var2csa in Pf3k Assembled Genomes

The locations of *var2csa* as determined by LASTZ(Harris 2007) in the chromosome level assemblies of the Pf3k genomes. **a)** PfTG01 is from Togo and has 4 copies have var2csa, two in tandem on chromosome 12 and two in tandem on chromosome 8. How related the copies are shown to the right. **b)** PfSN01 is from Senegal and has two var2csa copies in tandem on chromosome 12. **c)** Lab PfHB3 strain is from Honduras and has a var2csa copy on chromosome 12 and one on chromosome 1.

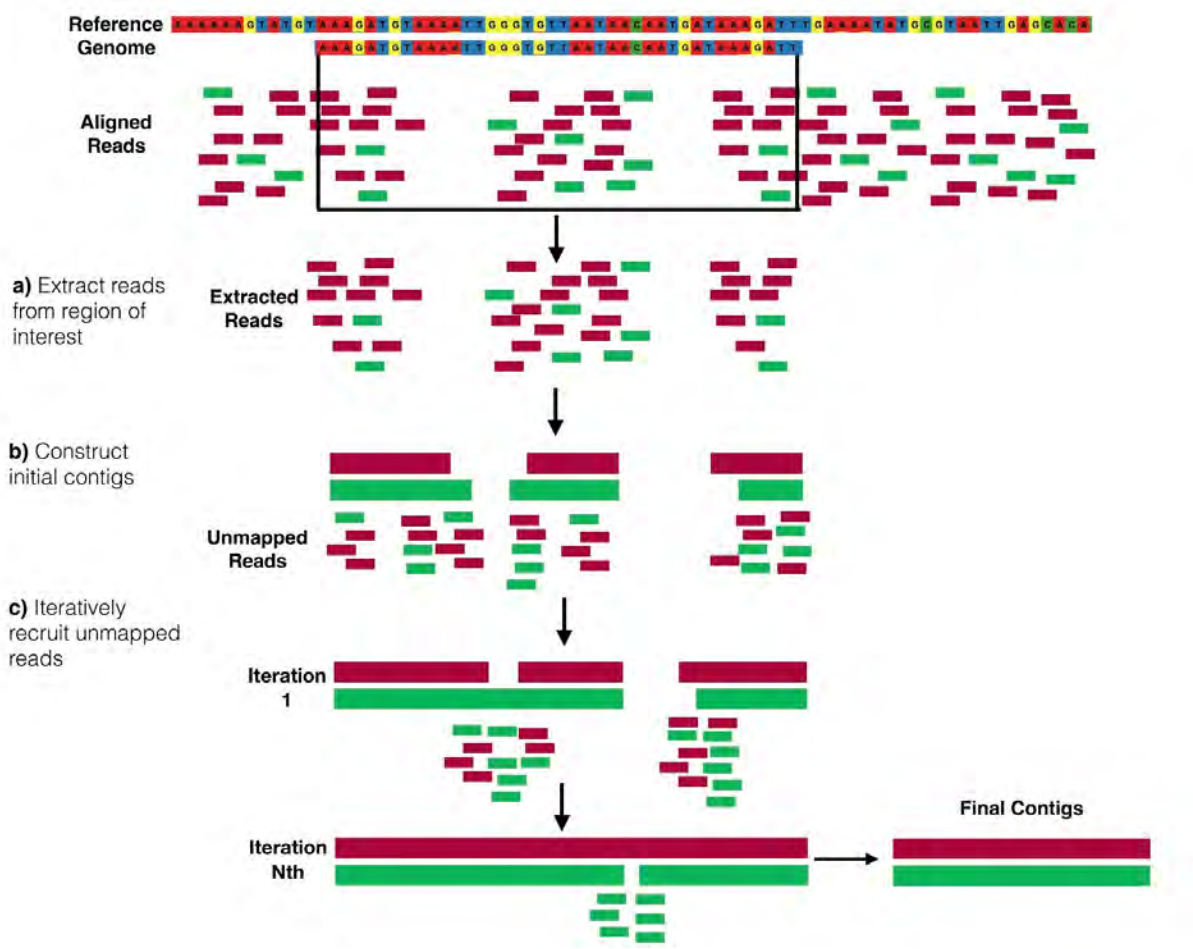


Figure 4.14: PathWeaver Recruitment Algorithm Overview

a) First what reads do map to a region are extracted and are used to construct initial contigs. **b)** These contigs are then used to recruit reads from the reads that are unmapped and assembly is ran again with all reads. **c)** This is done iteratively until there are either no more reads recruited or the max iterations (default 20) is hit.

Chapter V: Carmen: Where in the world is my haplotype?

Abstract

There have been many tools invented to view or summarize SNP/INDEL variant calls from shotgun whole genome sequencing; however, as targeted amplicon sequencing becomes more popular, tools that report haplotype information rather than the traditional variant calls will be more useful. For this reason, Carmen was invented to utilize the PathWeaver algorithm introduced in Chapter IV to both collect and visualize haplotype data from publicly available datasets to better inform targeted amplicon studies. Carmen was found to be accurate based off of results on various regions in the *P. falciparum* genome using known lab strain control mixture, and proved to be useful on a previous targeted amplicon dataset.

Introduction

As mentioned in previous chapters, the analysis of infectious disease by using haplotypes rather than just by calling SNP/INDEL variants is becoming more popular of late (Bailey et al. 2012; Mideo et al. 2016; R. H. Miller et al. 2017; Verity et al. 2018). However, there aren't as many programs that report haplotype worldwide prevalence like MalariaGen does with its Panoptes application for SNP variants (Vauterin et al. 2017). Haplotypes can also be checked in NCBI's BLAST to determine if a haplotype has been found before, but

this lacks the ability to report comprehensive summary reports on how many times, where, and when a haplotype has been found as appropriate metadata is often lacking or hard to collate. For that reason, I have invented the program dubbed Carmen to find where in the world certain haplotypes are found.

Carmen takes a genomic location, a directory of bam alignment files, and a metadata file for the input samples with country and collection year to collect local haplotypes for the given region and report the years in which each haplotype was found. Carmen was designed to be broadly applicable to any species, but was tested here on *P. falciparum* with several highly variable genomic regions, including regions often studied in targeted sequencing approaches, such as thrombospondin-related anonymous protein (TRAP), circumsporozoite protein (CSP) (Mideo et al. 2016; Bailey et al. 2012), and apical membrane antigen 1 (AMA1) (R. H. Miller et al. 2017), among many others; see **Table 5.1** and **Table 5.2** for a list of all genes. These regions were chosen because highly variable regions are the common target of targeted approaches and represent the most likely regions Carmen would be used on. Carmen was then tested for accuracy on known lab control mixtures provided by MalariaGen's Pf3k project (**Table 5.3**) and then used on output of a previous study on the CSP gene (Mideo et al. 2016).

Results

Known Lab Strains

The goal of Carmen is to collect as many high quality accurate local haplotype sequences from samples as possible, and not necessarily to call all expected sequences from a sample (especially since, depending on the length of the region of interest, this is not

always possible with a short read assembly method, as variation cannot always stitched together if read length is not adequate enough to bridge conserved sequence in between). For this reason, Carmen was evaluated for the number of haplotypes that matched the expected sequence, here coined “True haplotypes.” Results are summarized in Tables 5.4 and 5.5. Carmen shows high accuracy on all samples for all MOIs. For the 200 bp window, 20/31 (67%) of the samples showed perfect accuracy, with at most 5 windows out of the 1862 windows containing errors for a sample, and 1819/1862 (97.7%) of the windows were always reconstructed correctly. For the 400 bp window, 27/31 (87.1%) of the samples showed perfect accuracy, with at most 2 windows out of the 128 windows containing errors for a sample, and 123/128 (96.1%) of the windows were always reconstructed correctly. With the exception of the IT sample, Carmen was able to accurately reconstruct the majority of windows for both 200 bp and 400 bp for the monoclonal samples, showing that Carmen is able to accurately reconstruct all windows. The failure on the polyclonal samples are due to read length not being able to span the variation present in the multiple clones.

Example PfCSP Dataset

To test Carmen on a real set of haplotypes from a targeted amplicon analysis, I selected a previous study on the region of *P. falciparum* CSP encoding the polymorphic C-terminal region the gene, which resulted in 45 unique population haplotypes (Mideo et al. 2016). Carmen determined the genomic location of the 45 haplotypes to be Pf3D7_03_v3 221423-221670(-) in the 3D7 reference genome (v3), which is the correct region targeted in the study (247 bp long). Carmen used this determined region to call haplotypes from the samples described in Chapter IV which, in short, contain samples from the following countries: Bangladesh=50, Cambodia=663, DRC=113, French Guiana=58, Ghana=605,

Guinea=100, Laos=85, Malawi=269, Mali=119, Myanmar=60, Nigeria=5, Senegal=137, Thailand=536, The Gambia=65, Uganda=11, Vietnam=97, as well as the control datasets described above. The sequences extracted from the control dataset all matched the expected sequences for all 31 samples. Carmen was able to collect 3,181 (219 unique) sequences from the field samples.

Carmen was able to create a connected haplotype network by connecting all haplotypes that were 2 or fewer differences from each other (**Figure 5.1** PfCSP Network). The network is made up of nodes of haplotypes colored by the region found in, and the area of the circle corresponds to number of times it was found. It can be observed that there is some distinct clustering by geography, where haplotypes are more likely to cluster with other haplotypes found in the same region--which has been observed for this region previously (A. E. Barry et al. 2009). Nodes were also created for all 45 input population sequences, and colored by the country they came from in the previous study. The top 5 that appeared in the most subjects in the previous are labeled. We can see that the most abundant haplotype from the Cambodia samples (which was the dominant infection across all Cambodian subjects), matches perfectly with the most abundant haplotype from South East Asia; similarly, the most abundant haplotypes from the Tanzania samples cluster closely with haplotypes from Africa.

Discussion

As the focus moves from SNP/INDEL variant calling to targeted amplicon approaches, we will need more tools specifically for analyzing haplotype data. For that reason, I have created Carmen, a tool that can take advantage of the wealth of publicly available data which has associated country and collection year metadata to report

important prevalence data for newly found haplotypes. Carmen should be an important aid for interrupting targeted haplotype approaches.

I have evaluated Carmen's accuracy by utilizing control mixtures of known *P. falciparum* lab strains for which there are whole chromosome level assemblies available to check expected sequences against. I have demonstrated that Carmen can accurately reconstruct regions of interest of 200 bp and 400 bp which includes genes like AMA1 and CSP, which are often targets of targeted amplicon approaches in *P. falciparum* studies (R. H. Miller et al. 2017; Bailey et al. 2012; Mideo et al. 2016). Carmen has demonstrated the ability to accurately collect local haplotypes from the majority of windows from monoclonal samples, as well as from polyclonal samples, which allows it to optimize the amount of haplotype data it can collect and is not limited to only monoclonal samples.

Carmen was then used on a real dataset from a previous study (Mideo et al. 2016) using targeted amplicon sequencing of *P. falciparum* to create strain specific clearance curves for patients from Cambodia and Tanzania. Carmen was able to determine the appropriate region and extracted 3,181 (219 unique) from publically available field samples. Using the metadata associated with these field samples, we are able to see that the haplotypes from the Cambodian and Tanzanian samples matched haplotypes from the corresponding regions in the field samples. At the time of writing of the previous study (Mideo et al. 2016), there were some concerns whether to trust the data from the Cambodia dataset, since all of the samples were strongly dominated by a single haplotype; it was postulated that it could have been contamination. At that time, there wasn't a comprehensive way of determining whether this haplotype had been found before in Cambodia, other than by looking at haplotypes reported by other studies; however, now with Carmen, we can

confirm that the haplotype found is in fact the most dominant haplotype in the South East Asia region.

In conclusion, Carmen has proven to accurately extract haplotype sequence from publicly available shotgun whole genome field samples for *P. falciparum*; used on a real dataset, Carmen has shown the utility of being able to compare to expected haplotype sequence for both the genomic and geographical regions being studied. Carmen is written to be able to work with a variety of species, and should prove to be a valuable resource for targeted amplicon studies moving forward.

Methods

Algorithm Overview

Carmen can either be run by giving it a set of genomic location in a Browser Extensible Data (BED, https://genome.ucsc.edu/FAQ/FAQformat.html#format1) file, or a set of haplotypes from which Carmen will determine the genomic location by mapping to a reference genome using LASTZ (Harris 2007) to create a BED file. Carmen then runs the PathWeaver algorithm, described in Chapter IV, on the region(s) from the BED file on a directory of alignment of files. In short, the PathWeaver algorithm is a graph assembly based algorithm designed specifically to construct local haplotypes from a region of interest in an alignment file, with special care taken to avoid creating false haplotypes for the cases of multiple copies, or multiple clones for a region that share large amounts of conserved sequence. Carmen then collates the haplotypes called for a region from each sample into one file, and reports countries and years found in supplied with a metadata file, which is a tab separated file with one column being sample names and each additional column being a

meta field for the given samples. The only requirements of Carmen are that all alignment files are aligned to the same reference genome file; for inputs, Carmen also needs the reference genome that the samples were aligned to, and a metadata file that contains at least country and collection year (NAs can be provided as placeholders). In addition, other genomes or assemblies of known/lab strains can be provided, and the sequences of these strains will be added to the output. Carmen was designed to be broadly applicable and, therefore, as long as the input data matches the above requirements, Carmen should be able to run on input from a variety of sources.

Carmen's output is a directory which contains all the result files. These files include 1) a fasta file of collected sequences, given a unique identifier named with genomic location extracted from and appended with an ID number starting from 0 where 0 is the most commonly found haplotype, 1 is the second most commonly found haplotype, etc., 2) a report of the samples the haplotypes were found in and a summary of the years and countries these samples are from, and 3) if Carmen was run with input haplotypes rather than just a given BED file, a report of the newly constructed haplotypes that match or most closely match the input haplotypes. Carmen also comes with a lightweight HTML viewer that can be run on the output directory to interactively view where haplotypes are found on a map, an interactive sequence viewer, and a network of how all the haplotypes are related (**Figure 5.2**).

P. falciparum Known Control Mixtures

Though Carmen is designed to be broadly applicable, my lab primarily works with *P. falciparum*, and therefore its capabilities and accuracy was tested on input from *P. falciparum*. Carmen's accuracy was tested on 27 known lab strain control mixtures provided

by MalariaGen's Pf3k project (https://www.malariagen.net/projects/pf3k). The mixtures were of either 2 or 3 strains and consisted of the lab strains 3D7, Dd2, HB3, and 7G8; see Table 5.3 for details. Also, four additional monoclonal lab strains shotgun whole genome samples were analyzed, GB4 (ERR027100), 3D7 (ERR043381), IT (ERR713965), and Dd2 (ERR663287). These sample were chosen because Pf3k have assembled genomes for them and therefore expected sequences can be determined and checked for accuracy in reconstructions.

P. falciparum Genomic Locations Analyzed

Carmen was run on 1,843 windows of 200 base pairs (bp) in length, and 127 windows of 400 bp in length from an analysis on highly variable regions in *P. falciparum* (unpublished). The 200 bp windows covered 390 genes and the 400 bp windows covered 27 genes, see **Tables 5.1** and **5.2**. Window sizes of 200 bp and 400 bp were chosen as these are common sizes for targeted approaches due to sequence technology read length limitations of Illumina. Carmen was run on the lab control mixtures and monoclonal samples to extract local haplotypes from these regions. Extracted sequences compared to the expected sequence. Expected sequences were determined by extracting sequences from the Pf3k PacBio assembled genomes using LASTZ (Harris 2007) to determine their location. Data is only reported for a location if a haplotype that spans the whole region of interest is reconstructed. This is not always possible for polyclonal samples or for polycopy regions when the read length is insufficient to properly stitch together variation on the ends of conserved regions when the conserved region is longer than read length. In this case, segments that are shorter than the region of interest are created, but since the interest here is the local haplotype that spans the whole region, only full length local haplotypes are

collected. For this reason, some samples will have no local haplotypes called for a region which can happen more often as the number of strains/copies increases.

Example Dataset

To test Carmen, a real dataset a previous study had done in collaboration with our lab was used. The study had amplified the region of *P. falciparum* CSP encoding the polymorphic C-terminal region the gene in order to create strain specific clearance curves from samples from patients in Tanzania and Cambodia to look for evidence of slow clearing parasites (Mideo et al. 2016). The study had 14 patients sampled over 72 hours up to 4 times each. Analysis resulted in 45 final unique population haplotypes and these haplotypes were run through the Carmen pipeline.

Tables

Table 5.1: Gene IDs for 200 bp Windows

Gene ID	Gene Description	# of 200bp windows
PF3D7_0102500	erythrocyte binding antigen-181 (EBA181)	1
PF3D7_0102800	conserved Plasmodium protein, unknown function	2
PF3D7_0103100	vacuolar protein sorting-associated protein 51, putative (VPS51)	1
PF3D7_0103300	conserved Plasmodium protein, unknown function	5
PF3D7_0103500	conserved Plasmodium protein, unknown function	4
PF3D7_0103600	ATP-dependent RNA helicase, putative	2
PF3D7_0104100	conserved Plasmodium membrane protein, unknown function	7
PF3D7_0106300	calcium-transporting ATPase (ATP6)	4
PF3D7_0110200	FAD-linked sulfhydryl oxidase ERV1, putative (ERV1)	1
PF3D7_0110600	phosphatidylinositol-4-phosphate 5-kinase (PIP5K)	1
PF3D7_0112200	multidrug resistance-associated protein 1 (MRP1)	5
PF3D7_0112900	Plasmodium exported protein, unknown function	3
PF3D7_0113100	surface-associated interspersed protein 1.1 (SURFIN 1.1) (SURF1.1)	8
PF3D7_0113600	surface-associated interspersed protein 1.2 (SURFIN 1.2), pseudogene (SURF1.2)	11
PF3D7_0113800	DBL containing protein, unknown function	48
PF3D7_0202100	Plasmodium exported protein (PHISTc), unknown function,liver stage associated protein 2 (LSAP2)	9
PF3D7_0204500	aspartate aminotransferase,aspartate transaminase (AspAT)	8
PF3D7_0207000	merozoite surface protein 4 (MSP4)	1
PF3D7_0207300	serine repeat antigen 8 (SERA8)	3
PF3D7_0207500	serine repeat antigen 6 (SERA6)	3
PF3D7_0207700	serine repeat antigen 4 (SERA4)	1
PF3D7_0208000	serine repeat antigen 1 (SERA1)	5
PF3D7_0209000	6-cysteine protein (P230)	1
PF3D7_0210200	conserved Plasmodium protein, unknown function	2
PF3D7_0210800	conserved Plasmodium protein, unknown function	7
PF3D7_0211700	tyrosine kinase-like protein, putative (TKL1)	1
PF3D7_0212100	conserved Plasmodium protein, unknown function	4
PF3D7_0212400	conserved Plasmodium membrane protein, unknown function	4
PF3D7_0213800	conserved Plasmodium protein, unknown function	1
PF3D7_0214600	serine/threonine protein kinase, putative	2
PF3D7_0214800	conserved Plasmodium membrane protein, unknown function	1
PF3D7_0220000	liver stage antigen 3 (LSA3)	2
PF3D7_0220100	DnaJ protein, putative	9
PF3D7_0220800	cytoadherence linked asexual protein 2 (CLAG2)	1
PF3D7_0221100	Plasmodium exported protein, unknown function, pseudogene	1
PF3D7_0221200	Plasmodium exported protein (hyp15), unknown function	2
PF3D7_0301400	Plasmodium exported protein, unknown function	2
PF3D7_0301800	Plasmodium exported protein, unknown function	2

PF3D7_0301900	conserved Plasmodium protein, unknown function	2
PF3D7_0302100	serine/threonine protein kinase (SRPK1)	3
PF3D7_0302600	ABC transporter, (TAP family), putative	6
PF3D7_0303100	conserved Plasmodium protein, unknown function	1
PF3D7_0304600	circumsporozoite (CS) protein (CSP)	3
PF3D7_0304700	conserved Plasmodium protein, unknown function	1
PF3D7_0305100	conserved Plasmodium protein, unknown function	1
PF3D7_0308000	DNA polymerase epsilon subunit b, putative	1
PF3D7_0308100	conserved Plasmodium protein, unknown function	10
PF3D7_0309900	conserved Plasmodium protein, unknown function	5
PF3D7_0310200	phd finger protein, putative	1
PF3D7_0311600	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1, putative	1
PF3D7_0312100	E3 ubiquitin-protein ligase, putative	7
PF3D7_0315200	circumsporozoite- and TRAP-related protein (CTRP)	5
PF3D7_0315600	conserved Plasmodium protein, unknown function	6
PF3D7_0316200	conserved Plasmodium protein, unknown function	8
PF3D7_0318200	DNA-directed RNA polymerase II subunit RPB1, putative (RPB1)	7
PF3D7_0318300	conserved Plasmodium protein, unknown function	5
PF3D7_0319400	kinesin-8, putative	7
PF3D7_0319800	conserved Plasmodium protein, unknown function	1
PF3D7_0320400	oocyst capsule protein (Cap380)	12
PF3D7_0321500	peptidase, putative	3
PF3D7_0323400	Rh5 interacting protein (RIPR)	1
PF3D7_0401900	acyl-CoA synthetase (ACS6)	8
PF3D7_0402000	Plasmodium exported protein (PHISTa), unknown function	3
PF3D7_0402200	surface-associated interspersed protein 4.1 (SURFIN 4.1), pseudogene (SURF4.1)	22
PF3D7_0402300	reticulocyte binding protein homologue 1,normocyte binding protein 1 (RH1)	9
PF3D7_0402400	Plasmodium exported protein, unknown function (GEXP18)	14
PF3D7_0402800	erythrocyte membrane protein 1 (PfEMP1), pseudogene	6
PF3D7_0404600	conserved Plasmodium membrane protein, unknown function	5
PF3D7_0408600	sporozoite invasion-associated protein 1 (SIAP1)	23
PF3D7_0412300	phosphopantothenoylcysteine synthetase, putative	3
PF3D7_0413400	erythrocyte membrane protein 1 (PfEMP1), exon 1, pseudogene (VAR)	7
PF3D7_0414000	chromosome associated protein, putative	5
PF3D7_0414100	conserved Plasmodium membrane protein, unknown function	10
PF3D7_0414200.1	calmodulin-like+protein	4
PF3D7_0415200	conserved Plasmodium protein, unknown function	1
PF3D7_0415800	RING zinc finger protein, putative	1
PF3D7_0417200	bifunctional dihydrofolate reductase-thymidylate synthase (DHFR-TS)	2
PF3D7_0417400	conserved Plasmodium protein, unknown function	1
PF3D7_0418000	conserved Plasmodium protein, unknown function	12
PF3D7_0418300	conserved Plasmodium protein, unknown function	1
PF3D7_0418600	regulator of chromosome condensation, putative	7
PF3D7_0419000	conserved Plasmodium protein, unknown function	3
PF3D7_0419400	conserved Plasmodium protein, unknown function	1
PF3D7_0419900	phosphatidylinositol 4-kinase, putative	6
PF3D7_0420000	zinc finger protein, putative	12

PF3D7_0420100	serine/threonine protein kinase RIO2 (RIO2)	3
PF3D7_0420200	holo-(acyl-carrier protein) synthase, putative	9
PF3D7_0422200	erythrocyte+membrane-associated+antigen	8
PF3D7_0422500	pre-mRNA-splicing helicase BRR2, putative (BRR2)	1
PF3D7_0422800	serpentine receptor, putative (SR12)	3
PF3D7_0424100	reticulocyte binding protein homologue 5 (RH5)	7
PF3D7_0424300	erythrocyte binding antigen-165, pseudogene (EBA165)	5
PF3D7_0424400	surface-associated interspersed protein 4.2 (SURFIN 4.2) (SURF4.2)	29
PF3D7_0424600	Plasmodium exported protein (PHISTb), unknown function	7
PF3D7_0424900	Plasmodium exported protein (PHISTa), unknown function	3
PF3D7_0425000	Plasmodium exported protein, unknown function, pseudogene	1
PF3D7_0500900	serine/threonine protein kinase, FIKK family (FIKK5)	5
PF3D7_0501600	rhoptry-associated protein 2 (RAP2)	8
PF3D7_0501800	chromosome assembly factor 1 (CAF1)	3
PF3D7_0502400	ring-stage membrane protein 1,merozoite surface protein 8 (MSP8)	4
PF3D7_0503200	conserved Plasmodium protein, unknown function	1
PF3D7_0504700	conserved Plasmodium protein, unknown function	6
PF3D7_0506500	conserved Plasmodium protein, unknown function	3
PF3D7_0506900	rhomboid protease ROM4 (ROM4)	4
PF3D7_0508000	6-cysteine protein (P38)	8
PF3D7_0509600	asparagine--tRNA ligase (AsnRS)	7
PF3D7_0511400	conserved Plasmodium protein, unknown function	1
PF3D7_0514300	aspartate--tRNA ligase, putative	1
PF3D7_0515500	amino acid transporter, putative	2
PF3D7_0516100	cation-transporting ATPase 1 (ATPase1)	5
PF3D7_0518700	mRNA-binding protein PUF1 (PUF1)	1
PF3D7_0519300	cytochrome c oxidase assembly protein (heme A: farnesyltransferase), putative	1
PF3D7_0519900	conserved Plasmodium protein, unknown function	1
PF3D7_0522400	conserved Plasmodium protein, unknown function	3
PF3D7_0523000	multidrug resistance protein (MDR1)	6
PF3D7_0525100	acyl-CoA synthetase (ACS10)	15
PF3D7_0525800	inner membrane complex protein 1g, putative (IMC1g)	8
PF3D7_0526600	conserved Plasmodium protein, unknown function	2
PF3D7_0529000	conserved Plasmodium protein, unknown function	8
PF3D7_0529300	apicoplast TIC22 protein (TIC22)	2
PF3D7_0529400.1	conserved Plasmodium protein, unknown function	1
PF3D7_0529800	conserved Plasmodium protein, unknown function	1
PF3D7_0532200	Plasmodium exported protein (PHISTc), unknown function	1
PF3D7_0532300	Plasmodium exported protein (PHISTb), unknown function	13
PF3D7_0532600	Plasmodium exported protein, unknown function	2
PF3D7_0602400	elongation factor G (EF-G)	3
PF3D7_0602800	JmjC domain containing protein (JmjC2)	2
PF3D7_0603600	conserved Plasmodium protein, unknown function	9
PF3D7_0604100	transcription factor with AP2 domain(s),SPE2-interacting protein (SIP2)	5
PF3D7_0604300	conserved Plasmodium protein, unknown function	4
PF3D7_0606000	conserved Plasmodium protein, unknown function	5
PF3D7_0609600	probable protein, unknown function	4

PF3D7_0612700	6-cysteine protein (P12)	1
PF3D7_0612900	nucleolar GTP-binding protein 1, putative	6
PF3D7_0613300	rhoptry protein (ROP14)	6
PF3D7_0615400	ribonuclease, putative	5
PF3D7_0615900	conserved Plasmodium protein, unknown function	4
PF3D7_0619500	acyl-CoA synthetase (ACS12)	6
PF3D7_0619800	conserved Plasmodium membrane protein, unknown function	1
PF3D7_0620400	merozoite surface protein 10 (MSP10)	6
PF3D7_0622300	vacuolar transporter chaperone, putative	4
PF3D7_0622800	leucine--tRNA ligase, putative	1
PF3D7_0624800	conserved Plasmodium protein, unknown function	1
PF3D7_0625600	poly(A) polymerase PAP, putative	5
PF3D7_0625900	conserved Plasmodium protein, unknown function	2
PF3D7_0628100	HECT-domain (ubiquitin-transferase), putative	2
PF3D7_0628200	protein kinase PK4 (PK4)	1
PF3D7_0629700	SET domain protein, putative (SET1)	12
PF3D7_0630300	DNA polymerase epsilon, catalytic subunit a, putative	7
PF3D7_0630600	conserved Plasmodium protein, unknown function	3
PF3D7_0701900	Plasmodium exported protein, unknown function	7
PF3D7_0702000	Plasmodium exported protein (hyp12), unknown function	11
PF3D7_0702500	Plasmodium exported protein, unknown function	7
PF3D7_0703900	conserved Plasmodium membrane protein, unknown function	3
PF3D7_0704600	E3 ubiquitin-protein ligase (UT)	1
PF3D7_0705200	conserved Plasmodium protein, unknown function	2
PF3D7_0706100	conserved Plasmodium protein, unknown function	1
PF3D7_0709300	Cg2 protein (CG2)	8
PF3D7_0710000	conserved Plasmodium protein, unknown function	14
PF3D7_0710200	conserved Plasmodium protein, unknown function	4
PF3D7_0710900	50S ribosomal protein L1, mitochondrial, putative (RPL1)	3
PF3D7_0711200	conserved Plasmodium protein, unknown function	2
PF3D7_0713900	conserved Plasmodium protein, unknown function	2
PF3D7_0714200	conserved Plasmodium protein, unknown function	4
PF3D7_0716300	conserved Plasmodium protein, unknown function	5
PF3D7_0716700	conserved Plasmodium protein, unknown function	9
PF3D7_0716800	eukaryotic translation initiation factor 3 37.28 kDa subunit, putative	1
PF3D7_0721700	secreted ookinete protein, putative (PSOP1)	1
PF3D7_0723800	conserved Plasmodium protein, unknown function	1
PF3D7_0727000	vacuolar protein sorting-associated protein 53, putative (VPS53)	1
PF3D7_0727200	cysteine desulfurase, putative (NFS)	2
PF3D7_0728100	conserved Plasmodium membrane protein, unknown function	9
PF3D7_0729700	conserved Plasmodium protein, unknown function	2
PF3D7_0731400	serine/threonine protein kinase, FIKK family, pseudogene (FIKK7.2)	2
PF3D7_0731500	erythrocyte binding antigen-175 (EBA175)	26
PF3D7_0801300	von Willebrand factor A domain-related protein (WARP)	6
PF3D7_0802000	glutamate dehydrogenase, putative (GDH3)	6
PF3D7_0802300	rRNA processing WD-repeat protein, putative	4
PF3D7_0804500	conserved Plasmodium membrane protein, unknown function	6

PF3D7_0806200	conserved Plasmodium membrane protein, unknown function	2
PF3D7_0806300	ferlin, putative	4
PF3D7_0806500	DnaJ protein, putative	1
PF3D7_0806700	conserved Plasmodium membrane protein, unknown function	6
PF3D7_0808300	ubiquitin regulatory protein, putative	1
PF3D7_0809600	peptidase family C50, putative	2
PF3D7_0810600	RNA helicase, putative	7
PF3D7_0810800	dihydropteroate synthetase (DHPS)	6
PF3D7_0811300	CCR4-associated factor 1 (CAF1)	3
PF3D7_0811600	conserved Plasmodium protein, unknown function	3
PF3D7_0812900	probable protein, unknown function	1
PF3D7_0814600	conserved Plasmodium protein, unknown function	1
PF3D7_0816300	conserved Plasmodium protein, unknown function	1
PF3D7_0818800	U3 small nucleolar ribonucleoprotein protein, putative	1
PF3D7_0819400	perforin-like protein 4 (PLP4)	1
PF3D7_0820700	2-oxoglutarate+dehydrogenase+E1+component	2
PF3D7_0823800	DnaJ protein, putative	1
PF3D7_0825800	conserved Plasmodium protein, unknown function	6
PF3D7_0826500	ubiquitin conjugation factor E4 B, putative (UBE4B)	1
PF3D7_0826900	conserved Plasmodium protein, unknown function	2
PF3D7_0827100	translation initiation factor IF-2, putative	1
PF3D7_0827600	conserved Plasmodium protein, unknown function	3
PF3D7_0827800	SET domain protein, putative (SET3)	2
PF3D7_0827900	protein disulfide isomerase (PDI8)	3
PF3D7_0829000	conserved Plasmodium membrane protein, unknown function	1
PF3D7_0829600	early transcribed membrane protein 8 (ETRAMP8)	2
PF3D7_0829800	unspecified+product	1
PF3D7_0830100	unspecified+product	1
PF3D7_0830300	sporozoite invasion-associated protein 2 (SIAP2)	1
PF3D7_0830600	Plasmodium exported protein (PHISTc), unknown function	2
PF3D7_0830800	surface-associated interspersed protein 8.2 (SURFIN 8.2) (SURF8.2)	30
PF3D7_0831300	Plasmodium exported protein, unknown function (GEXP13)	7
PF3D7_0831400	Plasmodium exported protein, unknown function	6
PF3D7_0831600	cytoadherence linked asexual protein 8 (CLAG8)	1
PF3D7_0901700	Plasmodium exported protein (hyp5), unknown function	8
PF3D7_0902000	serine/threonine protein kinase, FIKK family (FIKK9.1)	3
PF3D7_0903300	conserved Plasmodium membrane protein, unknown function	1
PF3D7_0903400	DEAD/DEAH box helicase, putative	1
PF3D7_0903600.1	conserved Plasmodium protein, unknown function	2
PF3D7_0904300	conserved protein, unknown function	15
PF3D7_0905400	high molecular weight rhoptry protein 3 (RhopH3)	5
PF3D7_0905600	conserved Plasmodium protein, unknown function	4
PF3D7_0906400	dynein light intermediate chain 2, cytosolic	4
PF3D7_0910800	cytosolic Fe-S cluster assembly factor NBP35, putative (NBP35)	4
PF3D7_0911300	cysteine repeat modular protein 1 (CRMP1)	10
PF3D7_0912200	conserved Plasmodium membrane protein, unknown function	1
PF3D7_0912600	conserved Plasmodium protein, unknown function	2

PF3D7_0913900	arginine--tRNA ligase, putative	3
PF3D7_0914000	pseudouridylate synthase, putative	2
PF3D7_0914100	conserved Plasmodium protein, unknown function	1
PF3D7_0915400	6-phosphofructokinase (PFK9)	8
PF3D7_0916400	conserved Plasmodium protein, unknown function	1
PF3D7_0920200	CS domain protein, putative	2
PF3D7_0920700	conserved Plasmodium protein, unknown function	5
PF3D7_0922600	glutamine synthetase, putative	6
PF3D7_0924000	patatin-like phospholipase, putative	5
PF3D7_0926500	conserved Plasmodium protein, unknown function	2
PF3D7_0926600	conserved Plasmodium membrane protein, unknown function	11
PF3D7_0927200	zinc finger protein, putative	4
PF3D7_0929400	high molecular weight rhoptry protein 2 (RhopH2)	9
PF3D7_0930300	merozoite surface protein 1 (MSP1)	19
PF3D7_0935600	gametocytogenesis-implicated protein (GIG)	3
PF3D7_0936300	ring-exported protein 3 (REX3)	2
PF3D7_1001400	alpha/beta hydrolase, putative	3
PF3D7_1001600	alpha/beta hydrolase, putative	4
PF3D7_1001700	Plasmodium exported protein (PHISTc), unknown function	6
PF3D7_1002200	tryptophan-rich antigen 3 (PArt)	9
PF3D7_1004600	conserved Plasmodium membrane protein, unknown function	3
PF3D7_1005000	methionine--tRNA ligase, putative	2
PF3D7_1005300	conserved Plasmodium protein, unknown function	7
PF3D7_1005500	regulator of nonsense transcripts, putative	4
PF3D7_1009200	small subunit rRNA synthesis-associated protein, putative	1
PF3D7_1010800	50S ribosomal protein L22, mitochondrial, putative	4
PF3D7_1011500	conserved Plasmodium membrane protein, unknown function	2
PF3D7_1011800	PRE-binding protein (PREBP)	1
PF3D7_1013900	initiation factor 2 subunit family, putative	1
PF3D7_1016400	serine/threonine protein kinase, FIKK family (FIKK10.1)	5
PF3D7_1019300	zinc finger protein, putative	1
PF3D7_1020300	cytoplasmic dynein intermediate chain, putative	1
PF3D7_1022200	conserved Plasmodium membrane protein, unknown function	2
PF3D7_1023700	conserved Plasmodium protein, unknown function	8
PF3D7_1028500	conserved Plasmodium protein, unknown function	2
PF3D7_1029000	conserved Plasmodium protein, unknown function, pseudogene	1
PF3D7_1029100.1	conserved Plasmodium protein, unknown function	2
PF3D7_1029400	conserved Plasmodium protein, unknown function	4
PF3D7_1029600	adenosine deaminase (ADA)	5
PF3D7_1031400.1	OTU-like cysteine protease, putative	6
PF3D7_1032300	conserved Plasmodium protein, unknown function	1
PF3D7_1033000	conserved Plasmodium protein, unknown function	4
PF3D7_1034600	conserved Plasmodium protein, unknown function	3
PF3D7_1035000	U2 snRNA/tRNA pseudouridine synthase, putative	4
PF3D7_1035100	probable protein, unknown function	10
PF3D7_1035300	glutamate-rich protein (GLURP)	18
PF3D7_1035700	duffy binding-like merozoite surface protein (DBLMSP)	5

PF3D7_1036300	merozoite surface protein (DBLMSP2)	5
PF3D7_1036400	liver stage antigen 1 (LSA1)	11
PF3D7_1102400	flavoprotein, putative	7
PF3D7_1102500	Plasmodium exported protein (PHISTb), unknown function (GEXP02)	6
PF3D7_1102600	Plasmodium exported protein, unknown function (GEXP14)	2
PF3D7_1105600	translocon component PTEX88 (PTEX88)	5
PF3D7_1106800	protein kinase, putative	4
PF3D7_1111700	conserved Plasmodium protein, unknown function	2
PF3D7_1112100	conserved Plasmodium protein, unknown function	1
PF3D7_1113000	conserved Plasmodium protein, unknown function	12
PF3D7_1116800	heat shock protein 101,chaperone protein ClpB2 (HSP101)	2
PF3D7_1117200	conserved Plasmodium protein, unknown function	3
PF3D7_1118300	insulinase, putative	5
PF3D7_1120300	metal ion channel - Mg2+, Co2+ and Ni2+	10
PF3D7_1120400	alpha/beta hydrolase fold domain containing protein, putative	2
PF3D7_1121300	tyrosine kinase-like protein (TKL2)	5
PF3D7_1121800	petidase, M16 family	5
PF3D7_1125700	kelch protein, putative	1
PF3D7_1125800	kelch protein, putative	1
PF3D7_1126100	autophagy-related protein 7, putative (ATG7)	4
PF3D7_1126600	steryl ester hydrolase, putative	7
PF3D7_1128300	6-phosphofructokinase (PFK11)	5
PF3D7_1128900	conserved Plasmodium protein, unknown function	6
PF3D7_1129100	parasitophorous vacuolar protein 1 (PV1)	3
PF3D7_1131600	conserved Plasmodium protein, unknown function	2
PF3D7_1133400	apical membrane antigen 1 (AMA1)	29
PF3D7_1133900	conserved Plasmodium protein, unknown function	7
PF3D7_1135100	protein phosphatase 2C, putative	9
PF3D7_1139100	RNA-binding protein, putative	2
PF3D7_1140500	myosin F, putative (MyoF)	3
PF3D7_1140900	conserved Plasmodium protein, unknown function	4
PF3D7_1141000	conserved Plasmodium protein, unknown function	4
PF3D7_1141300	conserved Plasmodium protein, unknown function	1
PF3D7_1143500	conserved Plasmodium protein, unknown function	1
PF3D7_1143800	conserved Plasmodium protein, unknown function	3
PF3D7_1145200	serine/threonine protein kinase, putative	1
PF3D7_1145800	conserved Plasmodium protein, unknown function	10
PF3D7_1147500	farnesyltransferase beta subunit, putative	14
PF3D7_1200700	acyl-CoA synthetase (ACS7)	10
PF3D7_1201400	Plasmodium exported protein, unknown function	14
PF3D7_1205400	conserved Plasmodium protein, unknown function	7
PF3D7_1205900	conserved protein, unknown function	3
PF3D7_1208200	cysteine repeat modular protein 3 (CRMP3)	4
PF3D7_1208900	conserved Plasmodium protein, unknown function	2
PF3D7_1215900	serpentine receptor, putative (SR10)	1
PF3D7_1216600	cell traversal protein for ookinetes and sporozoites (CelTOS)	7
PF3D7_1217300	GTP-binding protein, putative	3

PF3D7_1217700	conserved Plasmodium protein, unknown function	3
PF3D7_1218000	thrombospondin-related apical membrane protein (TRAMP)	1
PF3D7_1218900	conserved Plasmodium protein, unknown function	6
PF3D7_1219000	formin+2	5
PF3D7_1219300	erythrocyte membrane protein 1, PfEMP1 (VAR)	2
PF3D7_1221000	histone-lysine N-methyltransferase, H3 lysine-4 specific (SET10)	1
PF3D7_1223400	phospholipid-transporting ATPase, putative	7
PF3D7_1226400	conserved Plasmodium protein, unknown function	3
PF3D7_1227500	cyclin (CYC2)	3
PF3D7_1228600	merozoite surface protein 9 (MSP9)	1
PF3D7_1228800	conserved Plasmodium protein, unknown function	8
PF3D7_1229500	T-complex protein 1, gamma subunit, putative	1
PF3D7_1230000	conserved Plasmodium protein, unknown function	1
PF3D7_1231000	conserved Plasmodium protein, unknown function	1
PF3D7_1231400	amino acid transporter, putative	5
PF3D7_1234200	conserved Plasmodium protein, unknown function, pseudogene	5
PF3D7_1235200	V-type K+-independent H+-translocating inorganic pyrophosphatase (VP2)	7
PF3D7_1235800	conserved Plasmodium protein, unknown function	2
PF3D7_1236400	conserved Plasmodium protein, unknown function	3
PF3D7_1237400	conserved Plasmodium protein, unknown function	2
PF3D7_1239800	conserved Plasmodium protein, unknown function	8
PF3D7_1239900	vacuolar protein sorting-associated protein 16, putative (VPS16)	4
PF3D7_1240200	erythrocyte membrane protein 1 (PfEMP1), pseudogene	16
PF3D7_1244400	conserved Plasmodium protein, unknown function	1
PF3D7_1244500	conserved Plasmodium protein, unknown function	6
PF3D7_1247500	serine/threonine protein kinase, putative	10
PF3D7_1248700	conserved Plasmodium protein, unknown function	4
PF3D7_1250100	osmiophilic body protein (G377)	2
PF3D7_1251200	coronin	8
PF3D7_1251700	tryptophan--tRNA ligase, putative,tryptophanyl-tRNA synthetase, putative (aTrpRS)	2
PF3D7_1252100	rhoptry neck protein 3 (RON3)	14
PF3D7_1252400	reticulocyte binding protein homologue 3, pseudogene (RH3)	9
PF3D7_1301600	erythrocyte binding antigen-140 (EBA140)	7
PF3D7_1301800	surface-associated interspersed protein 13.1 (SURFIN 13.1), pseudogene (SURF13.1)	5
PF3D7_1302900	conserved Plasmodium protein, unknown function	13
PF3D7_1303800	conserved Plasmodium protein, unknown function	7
PF3D7_1305000	conserved Plasmodium protein, unknown function	5
PF3D7_1306500	MORN repeat protein, putative	3
PF3D7_1308400	conserved Plasmodium protein, unknown function	12
PF3D7_1312600	2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial, putative (BCKDHA)	1
PF3D7_1312800	conserved Plasmodium protein, unknown function	3
PF3D7_1313600	conserved Plasmodium protein, unknown function	7
PF3D7_1318300	conserved Plasmodium protein, unknown function	6
PF3D7_1318900	conserved Plasmodium protein, unknown function	5
PF3D7_1320700	conserved Plasmodium protein, unknown function	15
PF3D7_1321100	conserved Plasmodium protein, unknown function	3
PF3D7_1321900	conserved Plasmodium protein, unknown function	3

PF3D7_1322300	translation+initiation+factor+EIF-2B+subunit+related	7
PF3D7_1327300	conserved Plasmodium protein, unknown function	6
PF3D7_1328200	conserved Plasmodium protein, unknown function	6
PF3D7_1331000	protein kinase, putative	8
PF3D7_1331500	conserved Plasmodium protein, unknown function	1
PF3D7_1333400	conserved protein, unknown function	3
PF3D7_1335100	merozoite surface protein 7 (MSP7)	18
PF3D7_1335900	sporozoite surface protein 2 (TRAP)	37
PF3D7_1340400	conserved Plasmodium protein, unknown function	3
PF3D7_1342600	myosin A (MyoA)	6
PF3D7_1342900	transcription factor with AP2 domain(s) (ApiAP2)	1
PF3D7_1343100	conserved Plasmodium protein, unknown function	2
PF3D7_1343400	DNA repair protein RAD5, putative (RAD5)	6
PF3D7_1343800	conserved Plasmodium protein, unknown function	2
PF3D7_1344000	aminomethyltransferase, putative	3
PF3D7_1344400	conserved Plasmodium protein, unknown function	5
PF3D7_1345600	inner+membrane+complex+protein	3
PF3D7_1346400	conserved Plasmodium protein, unknown function	2
PF3D7_1346700	6-cysteine protein (P48/45)	4
PF3D7_1347200	nucleoside transporter 1 (NT1)	2
PF3D7_1350500	conserved Plasmodium protein, unknown function	1
PF3D7_1352700	P-loop containing nucleoside triphosphate hydrolase, putative	3
PF3D7_1352900	Plasmodium exported protein, unknown function,fam-f protein	19
PF3D7_1353100	Plasmodium exported protein, unknown function	9
PF3D7_1355600	conserved Plasmodium protein, unknown function	2
PF3D7_1358200	conserved Plasmodium protein, unknown function	4
PF3D7_1358600	zinc finger protein, putative	2
PF3D7_1359500	conserved Plasmodium protein, unknown function	3
PF3D7_1359600	conserved Plasmodium protein, unknown function	3
PF3D7_1361800	conserved Plasmodium protein, unknown function	7
PF3D7_1362800	conserved Plasmodium protein, unknown function	1
PF3D7_1366300	conserved Plasmodium protein, unknown function	1
PF3D7_1366800	phosphatidylserine synthase, putative	2
PF3D7_1368800	DNA repair endonuclease, putative (ERCC4)	1
PF3D7_1369200	conserved Plasmodium protein, unknown function	3
PF3D7_1401200	Plasmodium exported protein, unknown function	1
PF3D7_1402100	conserved Plasmodium protein, unknown function	1
PF3D7_1403200	conserved Plasmodium protein, unknown function	1
PF3D7_1403300	conserved Plasmodium protein, unknown function	1
PF3D7_1404300	secreted ookinete adhesive protein (SOAP)	1
PF3D7_1406100	conserved Plasmodium protein, unknown function	4
PF3D7_1406500	conserved Plasmodium protein, unknown function	1
PF3D7_1406600	ATP-dependent Clp protease, putative (ClpC)	1
PF3D7_1407700	conserved Plasmodium protein, unknown function	8
PF3D7_1410300	conserved Plasmodium protein, unknown function	4
PF3D7_1410400	rhoptry-associated protein 1 (RAP1)	11
PF3D7_1412000	p1/s1 nuclease, putative	2

PF3D7_1414200	conserved Plasmodium protein, unknown function	3
PF3D7_1414900	ATP-dependent protease, putative	3
PF3D7_1415000	uracil-DNA glycosylase (UDG)	2
PF3D7_1415400	conserved Plasmodium protein, unknown function	3
PF3D7_1416100	root hair defective 3 GTP-binding protein (RHD3) homolog, putative	8
PF3D7_1416200	metacaspase-like protein (MCA3)	5
PF3D7_1417400	cyclic nucleotide-binding protein, putative, pseudogene (cNBP)	1
PF3D7_1417600	conserved Plasmodium protein, unknown function	5
PF3D7_1418100	liver specific protein 1, putative (LISP1)	2
PF3D7_1420100	conserved Plasmodium protein, unknown function	5
PF3D7_1422900	14-3-3 protein, putative	1
PF3D7_1423500	conserved Plasmodium protein, unknown function	1
PF3D7_1424400	60S ribosomal protein L7-3, putative	6
PF3D7_1426700	phosphoenolpyruvate carboxylase (PEPC)	9
PF3D7_1428500	protein kinase, putative	4
PF3D7_1429800	coatamer beta subunit, putative	3
PF3D7_1429900	ATP-dependent DNA helicase, putative	8
PF3D7_1434100	queuine tRNA-ribosyltransferase, putative	3
PF3D7_1436300	translocon component PTEX150 (PTEX150)	6
PF3D7_1440200	stromal-processing peptidase, putative (SPP)	2
PF3D7_1442200	GTP-binding protein, putative	3
PF3D7_1442400	conserved Plasmodium protein, unknown function	6
PF3D7_1442600	TRAP-like protein,sporozoite-specific transmembrane protein S6 (TREP)	4
PF3D7_1442700	conserved Plasmodium protein, unknown function	3
PF3D7_1443200	conserved Plasmodium protein, unknown function	4
PF3D7_1444100	conserved Plasmodium protein, unknown function	5
PF3D7_1445500	conserved Plasmodium protein, unknown function	1
PF3D7_1446500	conserved Plasmodium protein, unknown function	4
PF3D7_1447900	multidrug resistance protein 2 (heavy metal transport family) (MDR2)	6
PF3D7_1448200	conserved Plasmodium protein, unknown function	5
PF3D7_1448500	conserved Plasmodium protein, unknown function	7
PF3D7_1451600	LCCL domain-containing protein (LAP5)	3
PF3D7_1453900	conserved Plasmodium protein, unknown function	1
PF3D7_1454200	conserved Plasmodium protein, unknown function	2
PF3D7_1455800	LCCL domain-containing protein (CCp2)	3
PF3D7_1457400	conserved Plasmodium protein, unknown function	1
PF3D7_1457900	conserved Plasmodium protein, unknown function	5
PF3D7_1458300	conserved Plasmodium protein, unknown function	8
PF3D7_1460500	conserved Plasmodium protein, unknown function	3
PF3D7_1461800	conserved Plasmodium protein, unknown function	5
PF3D7_1462300	conserved Plasmodium protein, unknown function	2
PF3D7_1462400	conserved Plasmodium protein, unknown function	2
PF3D7_1464500	conserved Plasmodium membrane protein, unknown function	6
PF3D7_1465800	dynein beta chain, putative	9
PF3D7_1467600	conserved Plasmodium protein, unknown function	2
PF3D7_1467900	rab GTPase activator, putative	7
PF3D7_1469600	biotin carboxylase subunit of acetyl CoA carboxylase, putative (ACC)	4

PF3D7_1472200	histone deacetylase, putative (HDA1)	4
PF3D7_1472400	M1-family alanyl aminopeptidase, putative	2
PF3D7_1472700	DNA-directed RNA polymerase, alpha subunit, putative	1
PF3D7_1474000	probable protein, unknown function	2
PF3D7_1475100	conserved Plasmodium protein, unknown function	3
PF3D7_1475500	LCCL domain-containing protein (CCp1)	2
PF3D7_1475800	conserved Plasmodium protein, unknown function	12
PF3D7_1475900	conserved Plasmodium protein, unknown function	11
PF3D7_1476600	Plasmodium exported protein, unknown function	7
PF3D7_1477500	Plasmodium exported protein (PHISTb), unknown function	3
PF3D7_1477600	surface-associated interspersed protein 14.1 (SURFIN 14.1) (SURF14.1)	4
PF3D7_1478000	Plasmodium exported protein (PHISTa), unknown function (GEXP17)	8
PF3D7_1478600	EMP1-trafficking protein (PTP3)	5
PF3D7_1478700	Plasmodium exported protein, unknown function, pseudogene	3

Table 5.2: Gene IDs for 400 bp Windows

Gene ID	Gene Description	# of 400bp windows
PF3D7_0113800	DBL containing protein, unknown function	13
PF3D7_0202100	Plasmodium exported protein (PHISTc), unknown function,liver stage associated protein 2 (LSAP2)	1
PF3D7_0402200	surface-associated interspersed protein 4.1 (SURFIN 4.1), pseudogene (SURF4.1)	1
PF3D7_0402400	Plasmodium exported protein, unknown function (GEXP18)	3
PF3D7_0418000	conserved Plasmodium protein, unknown function	1
PF3D7_0420000	zinc finger protein, putative	4
PF3D7_0422200	erythrocyte+membrane-associated+antigen	2
PF3D7_0424400	surface-associated interspersed protein 4.2 (SURFIN 4.2) (SURF4.2)	12
PF3D7_0525100	acyl-CoA synthetase (ACS10)	5
PF3D7_0702000	Plasmodium exported protein (hyp12), unknown function	2
PF3D7_0731500	erythrocyte binding antigen-175 (EBA175)	12
PF3D7_0830800	surface-associated interspersed protein 8.2 (SURFIN 8.2) (SURF8.2)	8
PF3D7_0930300	merozoite surface protein 1 (MSP1)	5
PF3D7_1002200	tryptophan-rich antigen 3 (PArt)	1
PF3D7_1035100	probable protein, unknown function	1
PF3D7_1035300	glutamate-rich protein (GLURP)	3
PF3D7_1133400	apical membrane antigen 1 (AMA1)	18
PF3D7_1135100	protein phosphatase 2C, putative	4
PF3D7_1200700	acyl-CoA synthetase (ACS7)	1
PF3D7_1240200	erythrocyte membrane protein 1 (PfEMP1), pseudogene	3
PF3D7_1302900	conserved Plasmodium protein, unknown function	2
PF3D7_1320700	conserved Plasmodium protein, unknown function	1
PF3D7_1335100	merozoite surface protein 7 (MSP7)	5
PF3D7_1335900	sporozoite surface protein 2 (TRAP)	14
PF3D7_1352900	Plasmodium exported protein, unknown function,fam-f protein	1
PF3D7_1475800	conserved Plasmodium protein, unknown function	3
PF3D7_1475900	conserved Plasmodium protein, unknown function	1

Table 5.3: Known Lab Strain Control Mixture Percentages

sample	acc	3D7*	Dd2*	HB3*	7G8*
PG0389-C	ERS319116	90	10	0	0
PG0390-C	ERS319119	80	20	0	0
PG0391-C	ERS319122	67	33	0	0
PG0392-C	ERS319125	33	67	0	0
PG0393-C	ERS319128	20	80	0	0
PG0394-C	ERS319130	10	90	0	0
PG0395-C	ERS319132	0	33.3	33.3	33.3
PG0396-C	ERS319134	0	25	25	50
PG0397-C	ERS319136	0	14.3	14.3	71.4
PG0399-C	ERS319140	0	0	99	1
PG0400-C	ERS319142	0	0	95	5
PG0401-C	ERS319117	0	0	90	10
PG0402-C	ERS319120	0	0	85	15
PG0403-C	ERS319123	0	0	80	20
PG0404-C	ERS319126	0	0	75	25
PG0405-C	ERS319129	0	0	70	30
PG0406-C	ERS319131	0	0	60	40
PG0407-C	ERS319133	0	0	50	50
PG0408-C	ERS319135	0	0	40	60
PG0409-C	ERS319137	0	0	30	70
PG0410-C	ERS319139	0	0	25	75
PG0411-C	ERS319141	0	0	20	80
PG0412-C	ERS319143	0	0	15	85
PG0413-C	ERS319121	0	0	5	95
PG0414-C	ERS319124	0	0	1	99
PG0415-C	ERS319127	0	0	0	100
PG0398-C	ERS319138	0	0	100	0

* The relative abundance of each lab strain

Table 5.4: 200 bp Windows Results

sample	MOI*	True Haplotypes (%)	Total Haplotypes	# Windows Reconstructed Error Free (%)	# Windows Reconstructed
3D7	1	1862 (100%)	1862	1862 (100%)	1862
7G8	1	1862 (100%)	1862	1862 (100%)	1862
Dd2	1	1862 (100%)	1862	1862 (100%)	1862
GB4	1	1860 (100%)	1860	1860 (100%)	1860
HB3	1	1860 (100%)	1860	1860 (100%)	1860
IT	1	1769 (100%)	1769	1769 (100%)	1769
PG0389-C	2	2266 (99.7%)	2273	1622 (99.6%)	1629
PG0390-C	2	2796 (100%)	2796	1584 (100%)	1584
PG0391-C	2	2834 (100%)	2834	1602 (100%)	1602
PG0392-C	2	2828 (100%)	2828	1599 (100%)	1599
PG0393-C	2	2800 (100%)	2801	1598 (99.9%)	1599
PG0394-C	2	2142 (99.8%)	2146	1624 (99.8%)	1628
PG0399-C	2	1708 (100%)	1708	1704 (100%)	1704
PG0400-C	2	1637 (100%)	1637	1636 (100%)	1636
PG0401-C	2	2376 (99.9%)	2379	1655 (99.8%)	1658
PG0402-C	2	2819 (100%)	2819	1699 (100%)	1699
PG0403-C	2	2807 (100%)	2807	1666 (100%)	1666
PG0404-C	2	2853 (100%)	2853	1684 (100%)	1684
PG0405-C	2	2777 (100%)	2777	1646 (100%)	1646
PG0406-C	2	2795 (100%)	2795	1653 (100%)	1653
PG0407-C	2	2793 (100%)	2793	1652 (100%)	1652
PG0408-C	2	2766 (100%)	2766	1639 (100%)	1639
PG0409-C	2	2785 (100%)	2785	1648 (100%)	1648
PG0410-C	2	2828 (100%)	2829	1669 (99.9%)	1670
PG0411-C	2	2843 (100%)	2844	1677 (99.9%)	1678
PG0412-C	2	2582 (99.8%)	2586	1686 (99.8%)	1690
PG0413-C	2	1634 (99.8%)	1638	1608 (99.8%)	1612
PG0414-C	2	1853 (100%)	1853	1853 (100%)	1853
PG0395-C	3	3102 (100%)	3103	1401 (99.9%)	1402
PG0396-C	3	2989 (99.8%)	2994	1363 (99.6%)	1368
PG0397-C	3	2975 (99.5%)	2990	1425 (99%)	1439

*MOI=multiplicity of infection, number of strains in mixture

Table 5.5: 400 bp Windows Results

sample	MOI*	True Haplotypes (%)	Total Haplotypes	# Windows Reconstructed Error Free (%)	# Windows Reconstructed
3D7	1	128 (100%)	128	128 (100%)	128
7G8	1	128 (100%)	128	128 (100%)	128
Dd2	1	128 (100%)	128	128 (100%)	128
GB4	1	128 (100%)	128	128 (100%)	128
HB3	1	127 (100%)	127	127 (100%)	127
IT	1	116 (100%)	116	116 (100%)	116
PG0389-C	2	53 (100%)	53	39 (100%)	39
PG0390-C	2	99 (100%)	99	53 (100%)	53
PG0391-C	2	107 (100%)	107	57 (100%)	57
PG0392-C	2	99 (100%)	99	53 (100%)	53
PG0393-C	2	72 (100%)	72	40 (100%)	40
PG0394-C	2	74 (98.7%)	75	51 (98.1%)	52
PG0399-C	2	73 (100%)	73	73 (100%)	73
PG0400-C	2	79 (100%)	79	79 (100%)	79
PG0401-C	2	112 (98.2%)	114	68 (97.1%)	70
PG0402-C	2	128 (99.2%)	129	71 (98.6%)	72
PG0403-C	2	118 (100%)	118	65 (100%)	65
PG0404-C	2	123 (100%)	123	67 (100%)	67
PG0405-C	2	108 (100%)	108	60 (100%)	60
PG0406-C	2	129 (100%)	129	70 (100%)	70
PG0407-C	2	105 (100%)	105	58 (100%)	58
PG0408-C	2	109 (100%)	109	60 (100%)	60
PG0409-C	2	113 (100%)	113	62 (100%)	62
PG0410-C	2	129 (100%)	129	70 (100%)	70
PG0411-C	2	135 (100%)	135	73 (100%)	73
PG0412-C	2	105 (100%)	105	63 (100%)	63
PG0413-C	2	63 (100%)	63	63 (100%)	63
PG0414-C	2	128 (100%)	128	128 (100%)	128
PG0395-C	3	45 (100%)	45	19 (100%)	19
PG0396-C	3	46 (100%)	46	20 (100%)	20
PG0397-C	3	76 (97.4%)	78	36 (94.7%)	38

*MOI=multiplicity of infection, number of strains in mixture

Figures

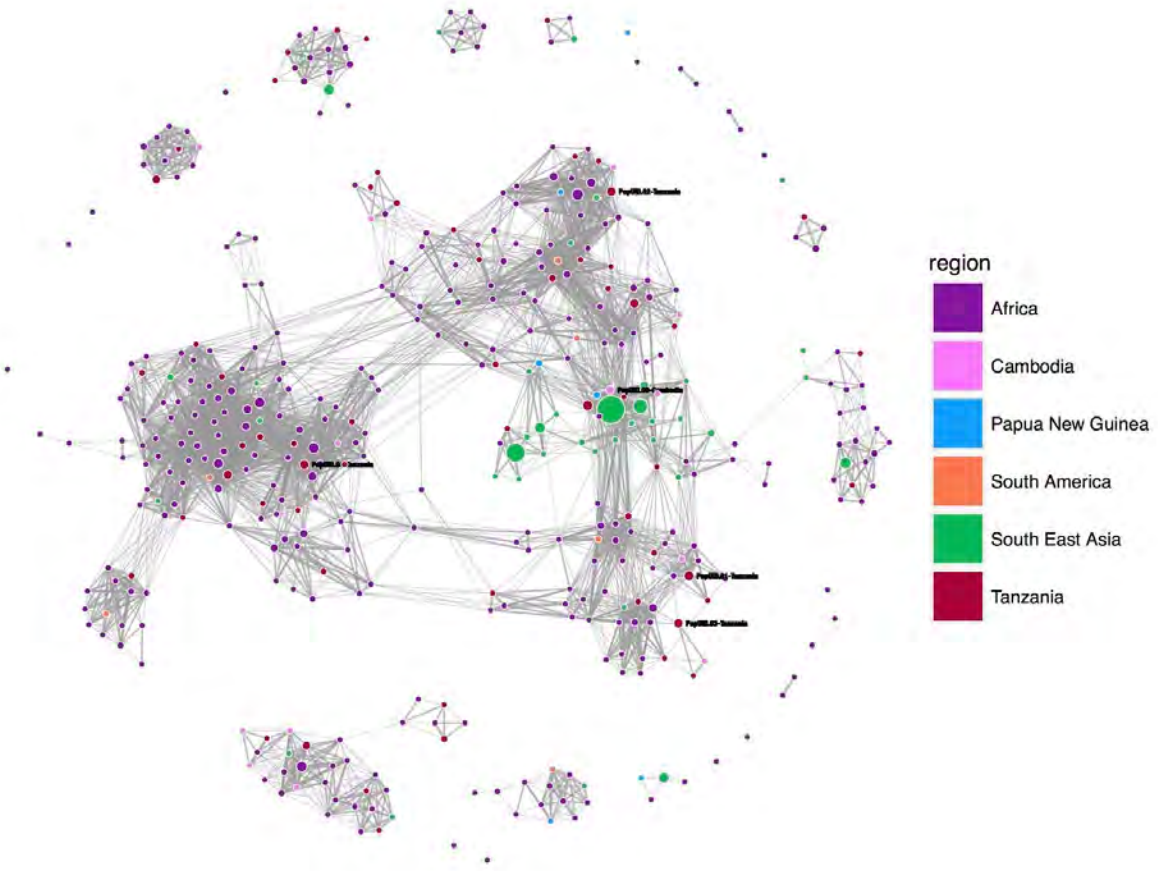


Figure 5.1: PfCSP Network

A network of PfCSP haplotypes extracted from publicly available data was created by generating nodes for haplotypes colored by the region they were found in and the area of the node is relative to the number of times a haplotype was found. Haplotypes are connected if they are 2 or fewer differences from each other. Nodes were also added from a previous study on CSP on patients from Tanzania and Cambodia which had resulted in 45 haplotypes and the top 5 haplotypes are labeled. The most abundant haplotype for the Cambodian samples can be clustered with the South East Asia haplotypes while the Tanzanian haplotypes are most clustered with the African samples.

Figure 5.2: Carmen Viewer Example

Carmen offers an interactive HTML viewer which can be used to view the sequences **a)** and offers some lightweight functionality like translating, running muscle, etc. **b)** An interactive map can be used to view where haplotypes appeared globally, when a haplotype's node is hovered over, all regions it was found in are highlighted for easy viewing.

Chapter VI: Discussion

In this thesis I have presented a suite of tools designed to analyze local haplotype-based approaches via high-throughput sequencing especially in the cases of polyclonal infections.

SeekDeep:

Chapter II introduced the SeekDeep pipeline which can be used to analyze targeted amplicon approaches using sequencing technologies 454, IonTorrent, and Illumina. SeekDeep has undergone heavy development over the years, with a focus on being as adaptive as possible. SeekDeep can handle several different technologies, and has default settings to handle each. SeekDeep was adapted to handle multiple different barcoding strategies, and can also handle doing a variable number of targets at once. One of the great strengths of SeekDeep is its ability to be able to recover data in even low read depths; this aids the recovery of data in hard to amplify samples and prevents the need to over-amplify samples which often leads to artifacts. The work on SeekDeep accumulated in its own publication (Hathaway et al. 2017) but has been used for a variety of studies for both *P. falciparum* (Mideo et al. 2016; R. H. Miller et al. 2017; Verity et al. 2018) and *P. vivax* (Lin et al. 2015) but can be--and is--used on a variety of infectious sources like the microbiome and HIV.

SeekDeep was used for a previous study on a region of *P. falciparum* CSP that encodes the polymorphic C-terminal region the gene. This study was with 8 patients from Tanzania and 8 from Cambodia and was done to detect the presence of slow clearing

parasites. The presence of slow clearing parasites has been linked to developing resistance to first line malaria treatment with Artemisinin based drug therapies (Noedl et al. 2008; Dondorp et al. 2009; Phyo et al. 2012). To detect the presence of slow clearing parasites, clearance curves are created by taking parasitemia levels following drug treatment. However, for complex infections with the presence of multiple different strains there could be a mix of drug-resistant and drug-susceptible parasites but a clearance curve based solely on parasitemia levels would represent an average clearance of all parasites and a slow clearing drug-resistance strain that was at low frequency within the infection would be masked due to the regular clearing of the major strains in the infections. For this reason, targeted amplicon sequencing was utilized to determine the relative frequencies of all strains present in the infection and strain specific clearance curves could be created. Up to 40% of the strains detected in the study differed from another strain by only a single base pair and due to decreasing parasitemia in the latter time points these samples often had low read depths of ~2,000. Therefore, in order to create accurate strain specific clearance curves SeekDeep's ability to detect single base difference between strains at various frequencies even at low read depths was essential, which is where SeekDeep has been able to outperform other programs (Hathaway et al. 2017).

Another study wherein SeekDeep's ability to detect single base differences was a great asset was a study conducted on *P. vivax* on the MSP1 gene (Lin et al. 2015). *P. vivax* has the ability to lay dormant in the liver of patients and disease can relapse after initial infection if dormant *P. vivax* parasite are released from the liver 3-4 weeks after drug treatment (White 2011). Relapse, which could be due to initial drug treatment failure, can be hard to distinguish from a new infection. In this study 78 adults were followed after an initial *P. vivax* infection and drug treatment for signs of recurrence of P. vivax in their blood.

Sequencing of the MSP1 gene was done for initial infection and for any follow up infections to determine relapse vs new infection. 70% of the parasites detected in the study were only a single base different from another parasite and therefore to adequately determine relapse vs new infection single base resolution was needed even if the parasite was at low frequency in order to determine the likelihood that specific parasites were in the initial infections, even if they were at low frequencies.

Though SeekDeep has primarily been used with malaria samples in the literature, it should also be useful in the study of other microbial populations that are also haploid-like viruses or bacterial populations.

kluster: Long Amplicon Clustering using k-mer Similarity Scores

SeekDeep's ability to analyze short amplicon sequences from technologies like Ion Torrent and Illumina was highlighted by Chapter II. Chapter III expands this work to longer amplicon approaches, primarily on PacBio sequencing which can be several kilobases in length. SeekDeep's core clustering algorithm, qluster, was ill suited to handle the high error rate of PacBio, and was further hampered by the longer amplicon length due to its dependence on alignment based comparison (which can slow exponentially with sequence length depending on implementation of the alignment approach). For this reason Chapter III introduces a novel clustering algorithm based on clustering sequences based on the k-mers (short sequence segments) shared by the sequences. This novel algorithm, kluster, can be used in place of the qluster algorithm from Chapter II and is integrated into the SeekDeep pipeline. This means the initial extraction and final population clustering can still be taken advantage of while using kluster. The performance of kluster proved to be as good as that of

qluster and can detect single base differences even with the increased sequence length and error rate of PacBio to read depths as low as 500 and frequencies of 1%.

The kluster algorithm has supported the study of *var2csa*, a protein expressed by *P. falciparum* that causes the parasite to sequester itself to placenta of pregnant woman leading to poor birth outcomes (Rogerson et al. 2007; Salanti et al. 2003a; Tuikue Ndam et al. 2005). Due to its high diversity, *var2csa* is hard to study via SNP variant calling; likewise, primers are difficult to design due to the target diversity, and the region responsible for binding is approximately 1.8kb long. These factors meant that the longer read lengths offered by PacBio were an ideal method to analyze the region. The kluster method was used to analyze the sequence of *var2csa* from pregnant woman from Benin and Malawi and to correlate these sequence birth outcomes. It was found that a specific clade of *var2csa* was associated with poor birth outcomes (Patel et al. 2017).

The kluster algorithm was then used on 15 of the women from Benin, comparing parasites found in the peripheral blood to parasites collected from the placenta. It was found that the same parasites found in the peripheral blood had the same *var2csa* haplotypes as the placental parasites for 13 out of the 15 women with the other 2 samples sharing the majority of the haplotypes from the two sites. This suggests that the parasites collected from the peripheral blood are a good approximation of what the parasites in the placenta, and the more invasive placental collection is not needed (Waltmann et al. 2018). As the goals of the study was to perfectly match up the haplotypes between the two different body sites it was critical that accurate sequence was called for each site to enable the matching of haplotypes and therefore the accuracy offered by kluster was essential for the study.

PathWeaver: Global Diversity of *P. falciparum var2csa*

While Chapter II and Chapter III introduced algorithms that worked on targeted amplicon sequencing, another common approach is to use shotgun whole genome sequencing, which generates reads starting from random locations across the whole genome. The targeted approach analysis is simplified by the fact all the reads start and end at the same location and clustering/analysis can be done by using all reads; however, the shotgun approach requires special handling because reads are from many different regions. For this reason, the traditional approach to analyzing shotgun whole genome sequencing is to map reads to a reference genome and call SNP/INDEL variants. While this approach works for much of the genome, for organisms like *P. falciparum* which has regions that are so diverse due to recombination or heavy immune selection that reads can't be mapped to the reference genome. These regions include key virulence factor genes in *P. falciparum* called *var* genes which encode EMP1, a protein that mediates the binding of infected erythrocytes to blood vessel walls often leading to the destruction of microvasculature and contribute to the more fatal clinical outcomes observed with *P. falciparum* infections like cerebral malaria. While several attempts have been made by utilizing genome assemblers like SPAdes(Lennartz et al. 2017; Jespersen et al. 2016), these studies did not check for accuracy of these assemblies and it was observed that these programs can lead to erroneous assemblies especially within mix infections (Chapter IV).

This dilemma led to the the development of the program PathWeaver which assembles local haplotypes for a given region of high diversity by first using the region to recruit initial reads to construct contigs by a graph assembly approach to then iteratively recruit unmapped sequences to these contigs until full local haplotypes are created.

PathWeaver was written with special care to not construct false haplotypes in multicopy/multiclonal scenarios where more than one unique sequence is present for a region, which can sometimes lead to variation belonging to separate copies being improperly stitched together, a common problem in graph assembly approaches. This was done by utilizing “threading” of sequences through the graph and making connections between variation only if supported by the underlying sequence data, an approach not utilized by SPAdes.

I was able to use the PathWeaver algorithm on a specific *var* gene of interest, *P. falciparum var2csa*, which causes poor birth outcomes in pregnant women due to its ability to bind to a placental protein CSA. With its high diversity, the PathWeaver algorithm was needed in order to properly collect genetic variation information for this gene due to only 80% of *var2csa* sequence being able to be mapped to reference. PathWeaver was extensively tested for accuracy and precision on the *var2csa* region using *in silico* simulations and monoclonal lab strains datasets where expected sequence was known. By utilizing approximately 3,000 field samples, PathWeaver was then used to collect between 1,000 and 2,000 sequences across the different domains of *var2csa* including the domain shown to be the domain most responsible for binding to CSA. Previous studies of *var2csa* only had approximately 30 sequences to analyze and so this number of collected sequences is a great improvement on fully elucidating the full global diversity of the gene.

PCAs on the minimum CSA binding domain showed 4 major groups and 2 minor groups that appear to be stable in the parasite population across geographical region and time which is suggestive of balancing selection (Lipsitch and O’Hagan 2007) acting on this region. Balancing selection is a phenomenon where diversity is maintained within a genomic region often due to due immune pressure and is often observed for surface epitopes

(Weedall and Conway 2010). This is especially true for a parasite like *P. falciparum* when people experience multiple infections and the increase in diversity aids the parasite in immune evasion when infecting a person not yet exposed to the parasite harboring specific epitopes. This causes a parasite's chance of survival to be inversely related to its frequency preventing fixation of a specific epitope. The evidence of balancing selection has implications for vaccine development for *var2csa* because, for a vaccine to protect against all parasites, it most likely will have to incorporate sequences from all the groups detected here. The two current vaccine trials only contain 1 strain each, 3D7 and FCR3, which both fall into different groups in the PCA and most likely will only protect against that group as even a single base difference has been shown to reduce the efficacy of a vaccine (Sedegah et al. 2016).

I was also able to extend previous findings of copy number variation (Sander et al. 2011, 2009) by finding multiple unique *var2csa* sequences across the whole gene within confirmed monoclonal samples. Each unique copy had the mean base genome coverage which supports that each copy is present in the genome. While previous studies have shown evidence of copy number variation, they were from only one country but here we were able to show that copy number variation is present in 21% of samples globally with up to 3 copies being observed in South East Asia and up to 5 copies in Africa though there was no evidence found in South America. This might be due to a sampling issue as there were only 34 samples from South America. The study of copy number variation would not have been possible without the ability of PathWeaver to be able to accurately assemble closely related copies of the same gene.

We have only scratched the surface for analyses that could be done here. However, conventional tools and metrics are not easily applied to a gene like *var2csa* with its complex

evolutionary history, high rate of recombination, high diversity preventing the ability to use one sequence as a reference and its multiple copies. Thanks to extremely divergent types there is no single good reference for sequences to be compared to which is the basis/requirement for many traditional measures of diversity and other population structure analyses. Also, a 3D structure could greatly inform the information gained from the sequence variation gathered here to see if variation is buried or forms pockets. There is currently no 3D structure available for *var2csa* but the amount of sequence gathered here could aid in the simulation of one.

This chapter was able to prove the accuracy of the PathWeaver algorithm as well as demonstrate the practical use of it on highly diverse *P. falciparum var2csa* which suggests it could be useful for other highly diverse regions of *P. falciparum* and other species.

Carmen: Where in the world is my haplotype?

As targeted approaches and other local haplotype-based analyses become more popular there will be a need for development for tools to view haplotype data similar to those created to view SNP data (Vauterin et al. 2017). For this purpose, Carmen was created. It utilizes the PathWeaver algorithm introduced in Chapter IV and the availability of thousands of publicly available shotgun whole genome sequence of field samples from around the world. While Carmen was written to be able to work for any input that can be aligned, it was primarily tested on highly diverse regions in *P. falciparum*. These regions were chosen as they are commonly targeted for amplicon approaches (R. H. Miller et al. 2017; Mideo et al. 2016; Bailey et al. 2012) and represent regions that Carmen is likely to be used on. Carmen was tested on monoclonal lab strains and mixtures of these lab strains for which there are whole genome assemblies available; this allowed checking against expected sequences.

Carmen was able to correctly reconstruct ~97% of the windows in all test samples and the majority of windows that had errors only had one or two false haplotypes from a single sample. Carmen had high accuracy (~99%) for windows reconstructed from samples with mixtures multiple strains, a scenario that may lead some assemblers to create false haplotypes.

Carmen takes advantage of the metadata associated with publicly available field samples to create summary reports of where and when haplotypes have been found for a specific genomic region. Carmen was written with the output of SeekDeep specifically in mind. For this reason, Carmen was used on the results from a previous study on *P. falciparum* CSP gene on patients from Tanzania and Cambodia to create strain specific clearance curves. Carmen was able to take the population haplotypes called from this study and accurately determine the appropriate genomic region by wrapping LASTZ (Harris 2007). Approximately 3,000 sequences were collected for this region and it was found that the most abundant haplotypes from Tanzania and Cambodia patients in the study matched the most prevalent haplotypes collected from the appropriate geographical regions. These results help mitigate past concerns that the haplotype from Cambodia could have been contamination, since it was found to be the dominant infection for all the Cambodian patients and demonstrate the utility of the results provided by Carmen.

Conclusion

In this thesis I have introduced a suite of tools for analyzing local haplotypes in complex mixtures from high-throughput sequencing with a focus on *Plasmodium falciparum* polyclonal infections. Though tested primarily on *Plasmodium*, the tools were written to take input general to the study of many different microbial populations and should be able to be

utilized for the study of populations similar to *Plasmodium* (e.g. viruses, other haploid organisms, bacterial populations, etc). This includes SeekDeep which analyzes targeted amplicon sequencing approaches on both short read technologies (like Ion Torrent and Illumina) and longer amplicon lengths with PacBio achieving one-base resolution. This one-base resolution enables the research of important targets for vaccine development or drug resistant genes, all of which can differ by only one base. I have also introduced the tool PathWeaver which has enabled an often ignored but very important virulence factor for malaria in pregnancy, *P. falciparum var2csa*. PathWeaver's analysis of this region has allowed for the further study of the global diversity and copy number variation to a much greater extent than previously possible which should greatly aid vaccine development. And lastly, Carmen uses the PathWeaver algorithm on the wealth of publicly available data to augment targeted amplicon analysis results by reporting on where and when haplotypes have been found before. These tools should hopefully prove to be useful to the field for years to come.

BIBLIOGRAPHY

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander. 2000. “An SNP Map of the Human Genome Generated by Reduced Representation Shotgun Sequencing.” *Nature* 407 (6803): 513–16.

Arez, Ana Paula, João Pinto, Katinka Pålsson, Georges Snounou, Thomas G. T. Jaenson, and Virgílio E. do Rosário. 2003. “Transmission of Mixed Plasmodium Species and Plasmodium Falciparum Genotypes.” *The American Journal of Tropical Medicine and Hygiene* 68 (2): 161–68.

Artyomenko, Alexander, Nicholas C. Wu, Serghei Mangul, Eleazar Eskin, Ren Sun, and Alex Zelikovsky. 2016. “Long Single-Molecule Reads Can Resolve the Complexity of the Influenza Virus Composed of Rare, Closely Related Mutant Variants.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, November. https://doi.org/10.1089/cmb.2016.0146.

Artyomenko, A., N. C. Wu, S. Mangul, E. Eskin, R. Sun, and A. Zelikovsky. 2015. “2SNV: Quasispecies Reconstruction from PacBio Reads.” In *2015 IEEE 5th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 1–1.

Ataíde, Ricardo, Alfredo Mayor, and Stephen J. Rogerson. 2014. “Malaria, Primigravidae, and Antibodies: Knowledge Gained and Future Perspectives.” *Trends in Parasitology* 30 (2): 85–94.

Bailey, Jeffrey A., Tisungane Mvalo, Nagesh Aragam, Matthew Weiser, Seth Congdon, Debbie Kamwendo, Francis Martinson, Irving Hoffman, Steven R. Meshnick, and Jonathan J. Juliano. 2012. “Use of Massively Parallel Pyrosequencing to Evaluate the Diversity of and Selection on Plasmodium Falciparum Csp T-Cell Epitopes in Lilongwe, Malawi.” *The Journal of Infectious Diseases* 206 (4): 580–87.

Baniecki, Mary Lynn, Aubrey L. Faust, Stephen F. Schaffner, Daniel J. Park, Kevin Galinsky, Rachel F. Daniels, Elizabeth Hamilton, et al. 2015. “Development of a Single Nucleotide Polymorphism Barcode to Genotype Plasmodium Vivax Infections.” *PLoS Neglected Tropical Diseases* 9 (3): e0003539.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77.

Barry, Alyssa E., Lee Schultz, Caroline O. Buckee, and John C. Reeder. 2009. “Contrasting Population Structures of the Genes Encoding Ten Leading Vaccine-Candidate Antigens of the Human Malaria Parasite, Plasmodium Falciparum.” *PloS One* 4 (12): e8497.

Barry, Alyssa, and Diana Hansen. 2016. “Naturally Acquired Immunity to Malaria.” *Parasitology* 143 (2): 125–28.

Basco, L. K., J. Le Bras, Z. Rhoades, and C. M. Wilson. 1995. “Analysis of pfmdr1 and Drug Susceptibility in Fresh Isolates of Plasmodium Falciparum from Subsaharan Africa.” *Molecular and Biochemical Parasitology* 74 (2): 157–66.

Beerenwinkel, Niko, Huldrych F. Günthard, Volker Roth, and Karin J. Metzner. 2012. "Challenges and Opportunities in Estimating Viral Genetic Diversity from next-Generation Sequencing Data." *Frontiers in Microbiology* 3 (September): 329.

Beerenwinkel, Niko, and Osvaldo Zagordi. 2011. "Ultra-Deep Sequencing for the Analysis of Viral Populations." *Current Opinion in Virology* 1 (5): 413–18.

Benítez-Páez, Alfonso, Kevin J. Portune, and Yolanda Sanz. 2016. "Species-Level Resolution of 16S rRNA Gene Amplicons Sequenced through the MinION™ Portable Nanopore Sequencer." *GigaScience* 5 (January): 4.

Bigey, Pascal, Sédami Gnidehou, Justin Doritchamou, Mickael Quiviger, Firmine Viwami, Aude Couturier, Ali Salanti, et al. 2011. "The NTS-DBL2X Region of VAR2CSA Induces Cross-Reactive Antibodies That Inhibit Adhesion of Several Plasmodium Falciparum Isolates to Chondroitin Sulfate A." *The Journal of Infectious Diseases* 204 (7): 1125–33.

Bordbar, Bita, Nicaise Tuikue-Ndam, Pascal Bigey, Justin Doritchamou, Daniel Scherman, and Philippe Deloron. 2012. "Identification of Id1-DBL2X of VAR2CSA as a Key Domain Inducing Highly Inhibitory and Cross-Reactive Antibodies." *Vaccine* 30 (7): 1343–48.

Bragg, Lauren, Glenn Stone, Michael Imelfort, Philip Hugenholtz, and Gene W. Tyson. 2012. "Fast, Accurate Error-Correction of Amplicon Pyrosequences Using Acacia." *Nature Methods* 9 (5): 425–26.

Brazeau, Nicholas F., Nicholas Hathaway, Christian M. Parobek, Jessica T. Lin, Jeffrey A. Bailey, Chanthap Lon, David L. Saunders, and Jonathan J. Juliano. 2016. "Longitudinal Pooled Deep Sequencing of the Plasmodium Vivax K12 Kelch Gene in Cambodia Reveals a Lack of Selection by Artemisinin." *The American Journal of Tropical Medicine and Hygiene*, October. https://doi.org/10.4269/ajtmh.16-0566.

Bucci, Vanni, Belinda Tzen, Ning Li, Matt Simmons, Takeshi Tanoue, Elijah Bogart, Luxue Deng, et al. 2016. "MDSINE: Microbial Dynamical Systems INference Engine for Microbiome Time-Series Analyses." *Genome Biology* 17 (1): 121.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods*, May. https://doi.org/10.1038/nmeth.3869.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. "QIIME Allows Analysis of High-Throughput Community Sequencing Data." *Nature Methods* 7 (5): 335–36.

Cerqueira, Gustavo C., Ian H. Cheeseman, Steve F. Schaffner, Shalini Nair, Marina McDew-White, Aung Pyae Phyo, Elizabeth A. Ashley, et al. 2017. "Longitudinal Genomic Surveillance of Plasmodium Falciparum Malaria Parasites Reveals Complex Genomic Architecture of Emerging Artemisinin Resistance." *Genome Biology* 18 (1): 78.

Cheeseman, Ian H., Becky Miller, John C. Tan, Asako Tan, Shalini Nair, Standwell C. Nkhoma, Marcos De Donato, et al. 2016. "Population Structure Shapes Copy Number Variation in Malaria Parasites." *Molecular Biology and Evolution* 33 (3): 603–20.

Chêne, Arnaud, Sophie Houard, Morten A. Nielsen, Sophia Hundt, Flavia D'Alessio, Sodiomon B. Sirima, Adrian J. F. Luty, et al. 2016. "Clinical Development of Placental Malaria Vaccines and Immunoassays Harmonization: A Workshop Report." *Malaria Journal* 15 (September): 476.

Clausen, Thomas M., Stig Christoffersen, Madeleine Dahlbäck, Annette Eva Langkilde, Kamilla E. Jensen, Mafalda Resende, Mette Ø. Agerbæk, et al. 2012. "Structural and Functional Insight into How the Plasmodium Falciparum VAR2CSA Protein Mediates Binding to Chondroitin Sulfate A in Placental Malaria." *The Journal of Biological*

Chemistry 287 (28): 23332–45.
Crosnier, Cécile, Zamin Iqbal, Ellen Knuepfer, Sorina Maciuca, Abigail J. Perrin, Gathoni Kamuyu, David Goulding, et al. 2016. “Binding of Plasmodium Falciparum Merozoite Surface Proteins DBLMSP and DBLMSP2 to Human Immunoglobulin M Is Conserved among Broadly Diverged Sequence Variants.” *The Journal of Biological Chemistry* 291 (27): 14285–99.
Dara, Antoine, Elliott F. Drábek, Mark A. Travassos, Kara A. Moser, Arthur L. Delcher, Qi Su, Timothy Hostelley, et al. 2017. “New Var Reconstruction Algorithm Exposes High Var Sequence Diversity in a Single Geographic Location in Mali.” *Genome Medicine* 9 (1): 30.
Dara, Antoine, Mark A. Travassos, Matthew Adams, Sarah Schaffer DeRoo, Elliott F. Drábek, Sonia Agrawal, Miriam K. Laufer, Christopher V. Plowe, and Joana C. Silva. 2017. “A New Method for Sequencing the Hypervariable Plasmodium Falciparum Gene var2csa from Clinical Samples.” *Malaria Journal* 16 (1): 343.
Dawson, Sarah-Jane, Dana W. Y. Tsui, Muhammed Murtaza, Heather Biggs, Oscar M. Rueda, Suet-Feung Chin, Mark J. Dunning, et al. 2013. “Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer.” *The New England Journal of Medicine* 368 (13): 1199–1209.
Dimonte, Sandra, Ellen I. Bruske, Johanna Hass, Christian Supan, Carmen L. Salazar, Jana Held, Serena Tschan, et al. 2016. “Sporozoite Route of Infection Influences In Vitro Var Gene Transcription of Plasmodium Falciparum Parasites From Controlled Human Infections.” *The Journal of Infectious Diseases* 214 (6): 884–94.
Dondorp, Arjen M., François Nosten, Poravuth Yi, Debashish Das, Aung Phae Phyo, Joel Tarning, Khin Maung Lwin, et al. 2009. “Artemisinin Resistance in Plasmodium Falciparum Malaria.” *The New England Journal of Medicine* 361 (5). Massachusetts Medical Society: 455–67.
Doritchamou, Justin, Audrey Sabbagh, Jakob S. Jespersen, Emmanuelle Renard, Ali Salanti, Morten A. Nielsen, Philippe Deloron, and Nicaise Tuikue Ndam. 2015. “Identification of a Major Dimorphic Region in the Functionally Critical N-Terminal ID1 Domain of VAR2CSA.” *PloS One* 10 (9): e0137695.
Duffy, Michael F., Alexander G. Maier, Timothy J. Byrne, Allison J. Marty, Salenna R. Elliott, Matthew T. O’Neill, Paul D. Payne, et al. 2006. “VAR2CSA Is the Principal Ligand for Chondroitin Sulfate A in Two Allogeneic Isolates of Plasmodium Falciparum.” *Molecular and Biochemical Parasitology* 148 (2): 117–24.
Duffy, Michael F., Jingyi Tang, Fransisca Sumardy, Hanh H. T. Nguyen, Shamista A. Selvarajah, Gabrielle A. Josling, Karen P. Day, Michaela Petter, and Graham V. Brown. 2017. “Activation and Clustering of a Plasmodium Falciparum Var Gene Are Affected by Subtelomeric Sequences.” *The FEBS Journal* 284 (2): 237–57.
Edgar, Robert C. 2013. “UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads.” *Nature Methods* 10 (10). Nature Publishing Group: 996–98.
———. 2016. “UNOISE2: Improved Error-Correction for Illumina 16S and ITS Amplicon Sequencing.” *bioRxiv*. https://doi.org/10.1101/081257.
Ester, M., H. P. Kriegel, J. Sander, and X. Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *KDD: Proceedings / International Conference on Knowledge Discovery & Data Mining. International Conference on Knowledge Discovery & Data Mining*. aaai.org. http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.
Evans, Andrew G., and Thomas E. Wellems. 2002. “Coevolutionary Genetics of

Plasmodium Malaria Parasites and Their Human Hosts." *Integrative and Comparative Biology* 42 (2): 401–7.

Friedman, Jonathan, and Eric J. Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." *PLoS Computational Biology* 8 (9): e1002687.

Fried, Michal, and Patrick E. Duffy. 2015. "Designing a VAR2CSA-Based Vaccine to Prevent Placental Malaria." *Vaccine* 33 (52): 7483–88.

Gardner, Malcolm J., Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W. Hyman, Jane M. Carlton, et al. 2002. "Genome Sequence of the Human Malaria Parasite Plasmodium Falciparum." *Nature* 419 (6906): 498–511.

Haas, Brian J., Dirk Gevers, Ashlee M. Earl, Mike Feldgarden, Doyle V. Ward, Georgia Giannoukos, Dawn Ciulla, et al. 2011. "Chimeric 16S rRNA Sequence Formation and Detection in Sanger and 454-Pyrosequenced PCR Amplicons." *Genome Research* 21 (3): 494–504.

Harris, R. S. 2007. "Improved Pairwise Alignment of Genomic DNA." The Pennsylvania State University. http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf.

Hathaway, Nicholas J., Christian M. Parobek, Jonathan J. Juliano, and Jeffrey A. Bailey. 2017. "SeekDeep: Single-Base Resolution de Novo Clustering for Amplicon Deep Sequencing." *Nucleic Acids Research*, November. https://doi.org/10.1093/nar/gkx1201.

Huang, Weichun, Leping Li, Jason R. Myers, and Gabor T. Marth. 2012. "ART: A next-Generation Sequencing Read Simulator." *Bioinformatics* 28 (4): 593–94.

Janda, J. Michael, and Sharon L. Abbott. 2007. "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls." *Journal of Clinical Microbiology* 45 (9): 2761–64.

Jayasundara, Duleepa, I. Saeed, Suhinthan Maheswararajah, B. C. Chang, S-L Tang, and Saman K. Halgamuge. 2015. "ViQuaS: An Improved Reconstruction Pipeline for Viral Quasispecies Spectra Generated by next-Generation Sequencing." *Bioinformatics* 31 (6): 886–96.

Jespersen, Jakob S., Christian W. Wang, Sixbert I. Mkumbaye, Daniel Tr Minja, Bent Petersen, Louise Turner, Jens Ev Petersen, John Pa Lusingu, Thor G. Theander, and Thomas Lavstsen. 2016. "Plasmodium Falciparum Var Genes Expressed in Children with Severe Malaria Encode CIDRα1 Domains." *EMBO Molecular Medicine* 8 (8): 839–50.

Juliano, Jonathan J., Kimberly Porter, Victor Mwapasa, Rithy Sem, William O. Rogers, Frédéric Ariey, Chansuda Wongsrichanalai, Andrew Read, and Steven R. Meshnick. 2010. "Exposing Malaria in-Host Diversity and Estimating Population Diversity by Capture-Recapture Using Massively Parallel Pyrosequencing." *Proceedings of the National Academy of Sciences of the United States of America* 107 (46): 20138–43.

Kembel, Steven W., Martin Wu, Jonathan A. Eisen, and Jessica L. Green. 2012. "Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance." Edited by Christian von Mering. https://doi.org/10.1371/journal.pcbi.1002743.

Kumar, Shiva, Devaraja G. Mudeppa, Ambika Sharma, Anjali Mascarenhas, Rashmi Dash, Ligia Pereira, Riaz Basha Shaik, et al. 2016. "Distinct Genomic Architecture of Plasmodium Falciparum Populations from South Asia." *Molecular and Biochemical Parasitology* 210 (1-2): 1–4.

Lakshmanan, Viswanathan, Patrick G. Bray, Dominik Verdier-Pinard, David J. Johnson, Paul Horrocks, Rebecca A. Muhle, George E. Alakpa, et al. 2005. "A Critical Role for PfCRT K76T in Plasmodium Falciparum Verapamil-Reversible Chloroquine

Resistance." *The EMBO Journal* 24 (13): 2294–2305.
Lavstsen, Thomas, Ali Salanti, Anja T. R. Jensen, David E. Arnot, and Thor G. Theander. 2003. "Sub-Grouping of Plasmodium Falciparum 3D7 Var Genes Based on Sequence Analysis of Coding and Non-Coding Regions." *Malaria Journal* 2 (September): 27.
Lennartz, Frank, Yvonne Adams, Anja Bengtsson, Rebecca W. Olsen, Louise Turner, Nicaise T. Ndam, Gertrude Ecklu-Mensah, et al. 2017. "Structure-Guided Identification of a Family of Dual Receptor-Binding PfEMP1 That Is Associated with Cerebral Malaria." *Cell Host & Microbe* 21 (3): 403–14.
Lerch, Anita, Cristian Koepfli, Natalie E. Hofmann, Camilla Messerli, Stephen Wilcox, Johanna H. Kattenberg, Inoni Betuela, Liam O'Connor, Ivo Mueller, and Ingrid Felger. 2017. "Development of Amplicon Deep Sequencing Markers and Data Analysis Pipeline for Genotyping Multi-Clonal Malaria Infections." *BMC Genomics* 18 (1): 864.
Lin, Jessica T., Nicholas J. Hathaway, David L. Saunders, Chanthap Lon, Sujata Balasubramanian, Oksana Kharabora, Panita Gosi, et al. 2015. "Using Amplicon Deep Sequencing to Detect Genetic Signatures of Plasmodium Vivax Relapse." *The Journal of Infectious Diseases* 212 (6): 999–1008.
Lipsitch, Marc, and Justin J. O'Hagan. 2007. "Patterns of Antigenic Diversity and the Mechanisms That Maintain Them." *Journal of the Royal Society, Interface / the Royal Society* 4 (16): 787–802.
Lysholm, Fredrik, Björn Andersson, and Bengt Persson. 2011. "An Efficient Simulator of 454 Data Using Configurable Statistical Models." *BMC Research Notes* 4 (October): 449.
MacIntyre, David A., Manju Chandiramani, Yun S. Lee, Lindsay Kindinger, Ann Smith, Nicos Angelopoulos, Benjamin Lehne, et al. 2015. "The Vaginal Microbiome during Pregnancy and the Postpartum Period in a European Population." *Scientific Reports* 5 (March): 8988.
Magoc, T., and S. L. Salzberg. 2011. "FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies." *Bioinformatics* 27 (21): 2957–63.
Mahé, Frédéric, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. 2014. "Swarm: Robust and Fast Clustering Method for Amplicon-Based Studies." *PeerJ* 2 (September): e593.
McColl, D. J., and R. F. Anders. 1997. "Conservation of Structural Motifs and Antigenic Diversity in the Plasmodium Falciparum Merozoite Surface Protein-3 (MSP-3)." *Molecular and Biochemical Parasitology* 90 (1): 21–31.
McManus, Kimberly F., Angela M. Taravella, Brenna M. Henn, Carlos D. Bustamante, Martin Sikora, and Omar E. Cornejo. 2017. "Population Genetic Analysis of the DARC Locus (Duffy) Reveals Adaptation from Standing Variation Associated with Malaria Resistance in Humans." *PLoS Genetics* 13 (3): e1006560.
Mideo, Nicole, Jeffrey A. Bailey, Nicholas J. Hathaway, Billy Ngasala, David L. Saunders, Chanthap Lon, Oksana Kharabora, et al. 2016. "A Deep Sequencing Tool for Partitioning Clearance Rates Following Antimalarial Treatment in Polyclonal Infections." *Evolution, Medicine, and Public Health* 2016 (1): 21–36.
Miller, Jason R., Sergey Koren, and Granger Sutton. 2010. "Assembly Algorithms for next-Generation Sequencing Data." *Genomics* 95 (6): 315–27.
Miller, Robin H., Nicholas J. Hathaway, Oksana Kharabora, Kashamuka Mwandagalirwa, Antoinette Tshefu, Steven R. Meshnick, Steve M. Taylor, Jonathan J. Juliano, V. Ann Stewart, and Jeffrey A. Bailey. 2017. "A Deep Sequencing Approach to Estimate Plasmodium Falciparum Complexity of Infection (COI) and Explore Apical Membrane Antigen 1 Diversity." *Malaria Journal* 16 (1): 490.

Miotto, Olivo, Roberto Amato, Elizabeth A. Ashley, Bronwyn MacInnis, Jacob Almagro-Garcia, Chanaki Amaratunga, Pharath Lim, et al. 2015. “Genetic Architecture of Artemisinin-Resistant Plasmodium Falciparum.” *Nature Genetics* 47 (3): 226–34.

Murat Eren, A., Hilary G. Morrison, Pamela J. Lescault, Julie Reveillaud, Joseph H. Vineis, and Mitchell L. Sogin. 2014. “Minimum Entropy Decomposition: Unsupervised Oligotyping for Sensitive Partitioning of High-Throughput Marker Gene Sequences.” *The ISME Journal* 9 (4). Nature Publishing Group: 968–79.

Nagesha, H. S., Din-Syafruddin, G. J. Casey, A. I. Susanti, D. J. Fryauff, J. C. Reeder, and A. F. Cowman. 2001. “Mutations in the pfmdr1, Dhfr and Dhps Genes of Plasmodium Falciparum Are Associated with in-Vivo Drug Resistance in West Papua, Indonesia.” *Transactions of the Royal Society of Tropical Medicine and Hygiene* 95 (1): 43–49.

Neafsey, Daniel E., Michal Juraska, Trevor Bedford, David Benkeser, Clarissa Valim, Allison Griggs, Marc Lievens, et al. 2015. “Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine.” *The New England Journal of Medicine* 373 (21): 2025–37.

Ngondi, Jeremiah M., Deus S. Ishengoma, Stephanie M. Doctor, Kyaw L. Thwai, Corinna Keeler, Sigsbert Mkude, Oresto M. Munishi, et al. 2017. “Surveillance for Sulfadoxine-Pyrimethamine Resistant Malaria Parasites in the Lake and Southern Zones, Tanzania, Using Pooling and next-Generation Sequencing.” *Malaria Journal* 16 (1): 236.

NIH HMP Working Group, Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A. Schloss, et al. 2009. “The NIH Human Microbiome Project.” *Genome Research* 19 (12): 2317–23.

Noedl, Harald, Youry Se, Kurt Schaecher, Bryan L. Smith, Duong Socheat, and Mark M. Fukuda. 2008. “Evidence of Artemisinin-Resistant Malaria in Western Cambodia.” *The New England Journal of Medicine* 359 (24). Massachusetts Medical Society: 2619–20.

Nwakanma, Davis C., Craig W. Duffy, Alfred Amambua-Ngwa, Eniyou C. Oriero, Kalifa A. Bojang, Margaret Pinder, Chris J. Drakeley, et al. 2014. “Changes in Malaria Parasite Drug Resistance in an Endemic Population over a 25-Year Period with Resulting Genomic Evidence of Selection.” *The Journal of Infectious Diseases* 209 (7): 1126–35.

Offeddu, Vittoria, Vandana Thathy, Kevin Marsh, and Kai Matuschewski. 2012. “Naturally Acquired Immune Responses against Plasmodium Falciparum Sporozoites and Liver Infection.” *International Journal for Parasitology* 42 (6): 535–48.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2018. “Vegan: Community Ecology Package.” https://CRAN.R-project.org/package=vegan.

Ouattara, Amed, Shannon Takala-Harrison, Mahamadou A. Thera, Drissa Coulibaly, Amadou Niangaly, Renion Saye, Youssouf Tolo, et al. 2013. “Molecular Basis of Allele-Specific Efficacy of a Blood-Stage Malaria Vaccine: Vaccine Development Implications.” *The Journal of Infectious Diseases* 207 (3): 511–19.

Parobek, Christian M., Jonathan B. Parr, Nicholas F. Brazeau, Chanthap Lon, Suwanna Chaorattanakawee, Panita Gosi, Eric J. Barnett, et al. 2017. “Partner-Drug Resistance and Population Substructuring of Artemisinin-Resistant Plasmodium Falciparum in Cambodia.” *Genome Biology and Evolution* 9 (6). Oxford University Press: 1673–86.

Patel, Jaymin C., Nicholas J. Hathaway, Christian M. Parobek, Kyaw L. Thwai, Mwayiwawo Madanitsa, Carole Khairallah, Linda Kalilani-Phiri, et al. 2017. “Increased Risk of Low Birth Weight in Women with Placental Malaria Associated with P. Falciparum VAR2CSA Clade.” *Scientific Reports* 7 (1): 7768.

Pearce, J. Andrew, Tony Triglia, Anthony N. Hodder, David C. Jackson, Alan F. Cowman, and Robin F. Anders. 2004. “Plasmodium Falciparum Merozoite Surface Protein 6 Is a Dimorphic Antigen.” *Infection and Immunity* 72 (4): 2321–28.

Phyo, Aung Pyae, Standwell Nkhoma, Kasia Stepniewska, Elizabeth A. Ashley, Shalini Nair, Rose McGready, Carit ler Moo, et al. 2012. “Emergence of Artemisinin-Resistant Malaria on the Western Border of Thailand: A Longitudinal Study.” *The Lancet* 379 (9830): 1960–66.

Prabhakaran, Sandhya, Melanie Rey, Osvaldo Zagordi, Niko Beerenwinkel, and Volker Roth. 2010. “HIV-Haplotype Inference Using a Constraint-Based Dirichlet Process Mixture Model.” In *Machine Learning in Computational Biology (MLCB) NIPS Workshop*, 1–4.

Quail, Michael A., Miriam Smith, Paul Coupland, Thomas D. Otto, Simon R. Harris, Thomas R. Connor, Anna Bertoni, Harold P. Swerdlow, and Yong Gu. 2012. “A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers.” *BMC Genomics* 13 (July): 341.

Quince, Christopher, Anders Lanzen, Russell J. Davenport, and Peter J. Turnbaugh. 2011. “Removing Noise from Pyrosequenced Amplicons.” *BMC Bioinformatics* 12 (January): 38.

Rask, Thomas S., Daniel A. Hansen, Thor G. Theander, Anders Gorm Pedersen, and Thomas Lavstsen. 2010. “Plasmodium Falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes--Divide and Conquer.” *PLoS Computational Biology* 6 (9). https://doi.org/10.1371/journal.pcbi.1000933.

Ricardo J G, Davoud Moulavi, and Joerg Sander. 2013. “Density-Based Clustering Based on Hierarchical Density Estimates.” In *Advances in Knowledge Discovery and Data Mining*, 160–72. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Rogerson, Stephen J., Lars Hviid, Patrick E. Duffy, Rose F. G. Leke, and Diane W. Taylor. 2007. “Malaria in Pregnancy: Pathogenesis and Immunity.” *The Lancet Infectious Diseases* 7 (2): 105–17.

Rowe, J. Alexandra, Antoine Claessens, Ruth A. Corrigan, and Mònica Arman. 2009. “Adhesion of Plasmodium Falciparum-Infected Erythrocytes to Human Cells: Molecular Mechanisms and Therapeutic Implications.” *Expert Reviews in Molecular Medicine* 11 (May): e16.

Salanti, Ali, Madeleine Dahlbäck, Louise Turner, Morten A. Nielsen, Lea Barfod, Pamela Magistrado, Anja T. R. Jensen, et al. 2004. “Evidence for the Involvement of VAR2CSA in Pregnancy-Associated Malaria.” *The Journal of Experimental Medicine* 200 (9): 1197–1203.

Salanti, Ali, Mafalda Resende, Sisse B. Ditlev, Vera V. Pinto, Madeleine Dahlbäck, Gorm Andersen, Tom Manczak, Thor G. Theander, and Morten A. Nielsen. 2010. “Several Domains from VAR2CSA Can Induce Plasmodium Falciparum Adhesion-Blocking Antibodies.” *Malaria Journal* 9 (January): 11.

Salanti, Ali, Trine Staalsoe, Thomas Lavstsen, Anja T. R. Jensen, M. P. Kordai Sowa, David E. Arnot, Lars Hviid, and Thor G. Theander. 2003a. “Selective Upregulation of a Single Distinctly Structured Var Gene in Chondroitin Sulphate A-Adhering Plasmodium Falciparum Involved in Pregnancy-Associated Malaria.” *Molecular Microbiology* 49 (1): 179–91.

———. 2003b. “Selective Upregulation of a Single Distinctly Structured Var Gene in Chondroitin Sulphate A-Adhering Plasmodium Falciparum Involved in Pregnancy-Associated Malaria.” *Molecular Microbiology* 49 (1): 179–91.

Salipante, Stephen J., Toana Kawashima, Christopher Rosenthal, Daniel R. Hoogestraat, Lisa A. Cummings, Dhruba J. Sengupta, Timothy T. Harkins, Brad T. Cookson, and Noah G. Hoffman. 2014. “Performance Comparison of Illumina and Ion Torrent next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling.” *Applied and Environmental Microbiology* 80 (24): 7583–91.

Sander, Adam F., Ali Salanti, Thomas Lavstsen, Morten A. Nielsen, Pamela Magistrado, John Lusingu, Nicaise Tuikue Ndam, and David E. Arnot. 2009. “Multiple var2csa-Type PfEMP1 Genes Located at Different Chromosomal Loci Occur in Many Plasmodium Falciparum Isolates.” *PloS One* 4 (8). Public Library of Science: e6667.

Sander, Adam F., Ali Salanti, Thomas Lavstsen, Morten A. Nielsen, Thor G. Theander, Rose G. F. Leke, Yeung Y. Lo, Naveen Bobbili, David E. Arnot, and Diane W. Taylor. 2011. “Positive Selection of Plasmodium Falciparum Parasites with Multiple var2csa-Type PfEMP1 Genes during the Course of Infection in Pregnant Women.” *The Journal of Infectious Diseases* 203 (11): 1679–85.

Schloss, Patrick D., Matthew L. Jenior, Charles C. Koumpouras, Sarah L. Westcott, and Sarah K. Highlander. 2016. “Sequencing 16S rRNA Gene Fragments Using the PacBio SMRT DNA Sequencing System.” *PeerJ* 4 (March). PeerJ Inc.: e1869.

Sedegah, Martha, Bjoern Peters, Michael R. Hollingdale, Harini D. Ganeshan, Jun Huang, Fouzia Farooq, Maria N. Belmonte, et al. 2016. “Vaccine Strain-Specificity of Protective HLA-Restricted Class 1 P. Falciparum Epitopes.” *PloS One* 11 (10): e0163026.

Seifert, David, Francesca Di Giallonardo, Armin Töpfer, Jochen Singer, Stefan Schmutz, Huldrych F. Günthard, Niko Beerenwinkel, and Karin J. Metzner. 2016. “A Comprehensive Analysis of Primer IDs to Study Heterogeneous HIV-1 Populations.” *Journal of Molecular Biology* 428 (1): 238–50.

Skums, Pavel, Zoya Dimitrova, David S. Campo, Gilberto Vaughan, Livia Rossi, Joseph C. Forbi, Jonny Yokosawa, Alex Zelikovsky, and Yury Khudyakov. 2012. “Efficient Error Correction for next-Generation Sequencing of Viral Amplicons.” *BMC Bioinformatics* 13 Suppl 10 (June): S6.

Smith, J. D., G. Subramanian, B. Gamain, D. I. Baruch, and L. H. Miller. 2000. “Classification of Adhesive Domains in the Plasmodium Falciparum Erythrocyte Membrane Protein 1 Family.” *Molecular and Biochemical Parasitology* 110 (2): 293–310.

Srivastava, Anand, Stéphane Gangnard, Sébastien Dechavanne, Farroudja Amirat, Anita Lewit Bentley, Graham A. Bentley, and Benoît Gamain. 2011. “Var2CSA Minimal CSA Binding Region Is Located within the N-Terminal Region.” *PloS One* 6 (5): e20270.

Taft, Diana H., Namasivayam Ambalavanan, Kurt R. Schibler, Zhuoteng Yu, David S. Newburg, Hitesh Deshmukh, Doyle V. Ward, and Ardythe L. Morrow. 2015. “Center Variation in Intestinal Microbiota Prior to Late-Onset Sepsis in Preterm Infants.” *PloS One* 10 (6): e0130604.

Taylor, Steve M., Jonathan J. Juliano, Paul A. Trottman, Jennifer B. Griffin, Sarah H. Landis, Paluku Kitsa, Antoinette K. Tshefu, and Steven R. Meshnick. 2010. “High-Throughput Pooling and Real-Time PCR-Based Strategy for Malaria Detection.” *Journal of Clinical Microbiology* 48 (2): 512–19.

Tuikue-Ndam, Nicaise, and Phillipe Deloron. 2015. “Developing Vaccines to Prevent Malaria in Pregnant Women.” *Expert Opinion on Biological Therapy* 15 (8). Taylor & Francis: 1173–82.

Tuikue Ndam, Nicaise G., Ali Salanti, Gwladys Bertin, Madeleine Dahlbäck, Nadine Fievet, Louise Turner, Alioune Gaye, Thor Theander, and Philippe Deloron. 2005. “High Level

of var2csa Transcription by Plasmodium Falciparum Isolated from the Placenta." *The Journal of Infectious Diseases* 192 (2): 331–35.

Vauterin, Paul, Ben Jeffery, Alistair Miles, Roberto Amato, Lee Hart, Ian Wright, and Dominic Kwiatkowski. 2017. "Panoptes: Web-Based Exploration of Large Scale Genome Variation Data." *Bioinformatics* 33 (20): 3243–49.

Verity, Robert, Nicholas J. Hathaway, Andreea Waltmann, Stephanie M. Doctor, Oliver J. Watson, Jaymin C. Patel, Kashamuka Mwandagalirwa, et al. 2018. "Plasmodium Falciparum Genetic Variation of var2csa in the Democratic Republic of the Congo." *Malaria Journal* 17 (1): 46.

Waltmann, Andreea, Jaymin C. Patel, Kyaw L. Thwai, Nicholas J. Hathaway, Christian M. Parobek, Achille Massougbodji, Nadine Fievet, et al. 2018. "Matched Placental and CirculatingPlasmodium falciparumParasites Are Genetically Homologous at thevar2csaID1-DBL2X Locus by Deep Sequencing." *The American Journal of Tropical Medicine and Hygiene* 98 (1): 77–82.

Wang, Bo, and Michael A. Kennedy. 2014. "Principal Components Analysis of Protein Sequence Clusters." *Journal of Structural and Functional Genomics* 15 (1): 1–11.

Ware, L. A., K. C. Kain, B. K. Lee Sim, J. D. Haynes, J. K. Baird, and D. E. Lanar. 1993. "Two Alleles of the 175-Kilodalton Plasmodium Falciparum Erythrocyte Binding Antigen." *Molecular and Biochemical Parasitology* 60 (1): 105–9.

Weedall, Gareth D., and David J. Conway. 2010. "Detecting Signatures of Balancing Selection to Identify Targets of Anti-Parasite Immunity." *Trends in Parasitology* 26 (7): 363–69.

White, Nicholas J. 2011. "Determinants of Relapse Periodicity in Plasmodium Vivax Malaria." *Malaria Journal* 10 (October): 297.

WHO. 2017. "World Malaria Report 2017." World Health Organization. http://www.who.int/malaria/publications/world-malaria-report-2017/report/en/.

Wirawan, Adrianto, Robert S. Harris, Yongchao Liu, Bertil Schmidt, and Jan Schröder. 2014. "HECTOR: A Parallel Multistage Homopolymer Spectrum Based Error Corrector for 454 Sequencing Data." *BMC Bioinformatics* 15 (May): 131.

Woo, P. C. Y., S. K. P. Lau, J. L. L. Teng, H. Tse, and K-Y Yuen. 2008. "Then and Now: Use of 16S rDNA Gene Sequencing for Bacterial Identification and Discovery of Novel Bacteria in Clinical Microbiology Laboratories." *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 14 (10): 908–34.

Yang, Xiao, Sriram P. Chockalingam, and Srinivas Aluru. 2013. "A Survey of Error-Correction Methods for next-Generation Sequencing." *Briefings in Bioinformatics* 14 (1): 56–66.

Zagordi, Osvaldo, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. 2011. "ShoRAH: Estimating the Genetic Diversity of a Mixed Sample from next-Generation Sequencing Data." *BMC Bioinformatics* 12 (April): 119.

Zhbannikov, Ilya Y., and James A. Foster. 2015. "MetAmp: Combining Amplicon Data from Multiple Markers for OTU Analysis." *Bioinformatics* 31 (11): 1830–32.