

Hypothesis-free detection of genome-changing events in pedigree sequencing



Kiran V Garimella
Green Templeton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2015

Dedication

Acknowledgements

Acknowledgements

Abstract

In high-diversity populations, a complete accounting of *de novo* mutations can be difficult to obtain. Most analyses involve alignment of genomic reads to a reference genome, but if the haplotypic background upon which a mutation occurs is absent, events can be easily missed (as reads have nowhere to align) and false-positives may abound (as the aligner forces the reads to align elsewhere). In this thesis, I describe methods for *de novo* mutation discovery and genotyping based on a so-called "pedigree graph" - a de Bruijn graph where all available sequencing data (trusted and untrusted data alike) is represented. I constrain genotyping efforts to locations containing "novel kmers" - sequence present in the child but absent from the parents. I then apply Dijkstra's shortest path algorithm to perform the genotyping, even in the presence of sequencing error. In simulation, this approach provides a vastly more sensitive and specific set of *de novo* variants than traditional methods.

In Chapter 1, I use part of a real dataset, progeny from the crossing of two *Plasmodium falciparum* parasites, to demonstrate the pitfalls of the reference-based approach. I also introduce de Bruijn graphs for genome assembly.

In Chapter 2, I present a review on mutational mechanisms that generate *de novo* mutations, their rates, factors that influence their generation, and known events in various species.

Chapters 3 and 4 detail the software packages I have written for this work, the former including descriptions of the realistic variant read simulations, the latter detailing the graph genotyping algorithm. Results from the application of the algorithm to the complete *P. falciparum* dataset are presented in Chapter 5.

Finally, Chapter 6 discusses the work in a larger context and details various improvements that can be made in future work.

Contents

1	Motivation	1
1.1	Introduction	1
1.2	<i>De novo</i> mutation (DNM)	2
1.2.1	Identification and evaluation of variation in genomes	3
1.2.2	A reference-free approach for assessing DNM sensitivity	4
1.3	DNM sensitivity of the reference-based approach	4
1.3.1	<i>Plasmodium falciparum</i>	4
1.3.1.1	Data processing	4
1.3.1.2	Results	8
1.4	<i>De novo</i> assembly as an alternative approach	9
2	Background	13
2.1	How genome changes	13
2.1.1	Cross-over	13
2.1.2	Gene conversion	13
2.1.3	Point mutations	13
2.1.4	Structural variants	13
2.1.4.1	Small (indels)	13
2.1.4.2	Large (fusions, NAHR)	13
2.1.4.3	Chromosomal changes	13
2.2	Rates	13
2.3	Factors influencing	13
2.3.1	Replication time	13
2.3.2	Mat/pat age effects	13
2.3.3	Biases/locality	13
2.4	Known events in species	13
2.4.1	P.f.	13
2.4.2	Human	13
2.4.3	Chimp	13

2.4.4	Others	13
3	Simulation	14
3.1	Simulating genomes	14
3.1.1	Parents	16
3.2	Children	19
3.2.1	Homologous recombination	19
3.2.1.1	Allelic	19
3.2.1.2	Gene conversion	19
3.2.1.3	Non-allelic	19
3.2.2	SNPs and indels	20
3.2.2.1	Expansion of short tandem repeats	22
3.2.2.2	Tandem duplications	22
3.3	Simulating reads	22
3.3.1	Constructing the coverage profile	23
3.3.1.1	ComputeBaseAndFragmentErrorRates	23
3.3.1.2	LiftoverFromRefToChild	24
3.3.2	Constructing the read profile	25
3.3.2.1	Scalar properties	25
3.3.2.2	Empirical distributions	26
3.3.2.3	Covariate table	26
3.3.3	The read simulator	27
4	Detection and genotyping	28
4.1	Introduction	28
4.2	Variant motifs	28
4.2.1	Simple variant motifs	28
4.2.2	Complex variant motifs	30
5	Pf	32
5.1	Lit review	32
5.1.1	Review of Kong et al., 2002	32
6	Chimp	33
6.1	Lit review	33
6.1.1	Review of Kong et al., 2002	33

7 Discussion	34
7.1 Lit review	34
7.1.1 Review of Kong et al., 2002	34
Bibliography	34

List of Figures

1.1	a. Parental and child sequences at the site of a <i>de novo</i> mutation, and the kmers generated at $k = 3$. b. The resulting Venn diagram of kmers found exclusively in the parents, the child, or common to both. c. Novel kmers found around the genome indicate the presence of a <i>de novo</i> mutation.	6
1.2	Kmer coverage distribution for a single 3D7xHB3 progeny, PG0063-C. Red line indicates non-parametric LOESS fit upon which the local minimum is detected. . . .	7
1.3	Novel kmers observed in the reference-based analysis vs trusted novel kmers expected from the reference-free analysis.	9
1.4	Reads that map once to the reference genome versus mapping multiple times, conditioned on the read containing a trusted novel kmer.	10
1.5	Venn diagram of kmers present in the 3D7 and HB3 genomes at $k = 47$. Both forward and reverse-complement kmers are considered the same.	11
1.6	Presence and absence of unique kmers in three 3D7 <i>var</i> genes. Each vertical line represents a kmer. Colored kmers represent those unique 3D7 kmers that are recovered in the sample. Grey indicates no recovery. White indicates the kmer at that position was not unique in the 3D7 genome. Only the coding regions of the respective genes are shown, with domain annotations obtained from the VarDom server. ¹	12
3.1	Workflow for generating the HB3 parental genome. a. Reads from HB3 sample, PG0052-C, are mapped to the 3D7 reference genome, and variants (SNPs and indels) are called and stored as a VCF file. b. We remove the 3D7 <i>var</i> gene repertoire, replacing each with a reasonable HB3 <i>var</i> counterpart, and encode the changes to the 3D7 reference genome as a VCF file. c. We alter the reference genome using Algorithm 2, thus producing the simulated HB3 genome.	17
3.2	Empirical recombination frequencies per chromosome	20
3.3	Simulated haplotype mosaics for chromosome 12.	21
3.4	Non-allelic recombinations for two compatible <i>var</i> genes.	21
3.5	Expansion and contraction of short tandem repeats (STRs)	22

4.1	a. Haploid sequences from a mother (green), father (blue), and child (red), the last differing from the first two by a single SNP. b. The resulting multi-color de Bruijn graph for $k = 3$. Red vertices denote kmers that are deemed "novel", i.e. present in the child and absent in the parents. Edge colors reflect the samples in which the connected pairs of kmers are found. Edges that are part of the bubble (variant call) are displayed with thicker lines.	29
4.2	A multi-color de Bruijn graph at $k = 47$ for a haploid pedigree spanning a simulated <i>de novo</i> SNP. Vertex labels have been suppressed for clarity. Spatial layout is arbitrary and for display purposes only.	29
4.3	A 5 bp insertion in the child	30
4.4	A 5 bp deletion in the child	30
4.5	A tandem duplication on the haplotypic background of the mother.	31
4.6	A variant wherein the child's path does not simply diverge from that of the parents, but rather navigates both.	31

List of Tables

1.1	Phenotypes of <i>P. falciparum</i> isolates used for genetic crosses.	2
1.2	Theoretical percentage of the genome recovered at a target depth of coverage. . . .	5
1.3	The <i>P. falciparum</i> datasets that will be referred to throughout this work.	5
3.1	Assembly statistics on publicly available finished and draft <i>P. falciparum</i> references, ordered by scaffold N50 length. Parental samples are shown in boldface.	16
3.2	Variants found the HB3 (PG0052-C) sample from the MalariaGen 3D7xHB3 dataset.	17
3.3	<i>Var</i> gene replacements from 3D7 to HB3 repertoire.	18

List of Algorithms

1	Given a set of variants, generate all possible subsets of variants	8
2	Generate an alternative reference sequence based on a VCF file.	15
3	Emit all fragment starts, read starts, and error rates per position.	24

1 Motivation

1.1 Introduction

IN 1905, WILLIAM BATESON, EDITH SAUNDERS, AND REGINALD PUNNETT published results on their work with dihybrid crosses of pea plants. Having previously established the dominance of two phenotypes (purple flowers over red, and long pollen grains over round), the trio crossed pea plants exhibiting both dominant or both recessive traits for successive generations. Crossing the first generation was expected to yield progeny with phenotype ratios following the 9:3:3:1 pattern according to the Law of Independent Assortment, established by Gregor Mendel 40 years prior. However, the observed ratios differed wildly from this expectation, heavily skewed towards the parental combination of phenotypes. Bateson *et al.* hypothesized that alleles controlling the two traits must be linked and transmitted to the progeny together by some unknown mechanism.²

Seven years later, Thomas Morgan noted an even more aberrant inheritance pattern in a dihybrid cross of white-eyed male and red-eyed female *Drosophila melanogaster* flies: no white-eyed females were observed in the progeny. A follow-up experiment crossing the original white-eyed male with its F_1 daughters produced four roughly equal-sized groups of red-eyed and white-eyed males and females, suggesting the white-eyed factor must be linked to the male sex factor. Morgan proposed that these factors (what we now refer to as genes) are real entities, physically located on chromosomes, and when found on the same chromosome, do not (necessarily) sort independently.³

Today, linkage analysis in experimentally controlled crosses (like the pea plants and fruit flies) or pedigrees (multi-generational families) is an invaluable tool in the study of genetics. Progeny only inherit half of their genome from each parent, so having several progeny that still exhibit the parental phenotypes helps narrow the region of the genome one must search to find the relevant genes. Modern experiments exploit the recombinant nature of genomes to unravel myriad phenotypes, the genes that control them, and the underlying genome biology responsible. These approaches have been broadly applied, proving particularly valuable in pathogens. For example, crosses of *Plasmodium falciparum* parasites, the causative agent for the most deadly form of malaria, have led to the discovery of a number of virulence factors. Several contrasting phenotypes for

Table 1.1: Phenotypes of *P. falciparum* isolates used for genetic crosses.

	3D7	HB3	DD2	7G8	GB4	803
pyrimethamine sensitivity	-	+				
chloroquine sensitivity		+	-			
infects mosquitoes easily		+	-			
infects <i>Aotus nancymae</i>				-	+	-
artemisinin sensitivity					+	-

various strains of *P. falciparum* are listed in Table 1.1. Crossing of the pyrimethamine-resistant 3D7 and sensitive HB3 strains⁴ revealed a nonsynonymous point mutation in the *dhfr-ts*¹ gene, inhibiting binding of (and thus conferring resistance to) the drug.⁵ Analysis of the HB3 x DD2 cross,⁶ the latter of which is resistant to chloroquine, localized the determinant to a previously undetected gene on chromosome 7, labelled *pfcr2*, a member of a new family of transporters. Additional investigation into differences in mosquito infection efficacy revealed down-regulation of the *pfmdv-1*³ gene,^{7,8} disruption of which results in marked reduction of mature and functional male gametocytes.

1.2 *De novo* mutation (DNM)

Many clinically relevant phenotypes are not inherited, but instead arise anew within the children. Cytoadherence and antigenic properties of parasites facilitate evasion from host immune attack, and can differ substantially from the properties of their progenitors. In 2000, Freitas-Junior *et al.* showed that some 3D7 x HB3, HB3 x DD2, and HB3 x HB3 progeny harbored non-parental forms of subtelomeric *var* genes, key members of an antigenic gene family.⁹ These altered forms were likely generated during mitosis by non-allelic homologous recombination⁴ (NAHR) of telomeres from two different chromosomes.¹⁰ The resulting genes are novel, functional, and never before observed by the host immune system.

These uninherited and spontaneous, or "*de novo*" mutations (DNMs) are critical tools for pathogenic adaption (and perhaps in higher-order organisms, maladaptation). Pathogens are under constant immune and intermittent drug pressure. NAHR can further diversify a pathogen's antigenic repertoire, enabling continued evasion of immunological actors. Duplications of a transporter gene may enable faster drug clearance in a parasite, thus conferring resistance. Spontaneous point mutations may alter the binding site of a drug to a receptor, thus conferring immunity.

¹ dihydrofolate reductase-thymidylate synthase

² *P. falciparum* chloroquine resistance transporter

³ *P. falciparum* male gametocyte development gene 1

⁴ sometimes referred to as "ectopic" (aberrant) recombination in the literature

1.2.1 Identification and evaluation of variation in genomes

With the advent of sequencing technologies, it is now straightforward to discover many DNMs. Long reads (~500 bp) from the first-generation sequencing technology, Sanger sequencing, can be stitched together *in silico* by considering the overlaps of sequences generated from many copies of the genome.¹¹ This has enabled the assembly of full-length genomes and subsequent gene annotations for a single representative (or "reference") individual in a population. Second-generation sequencing is leveraging economies of scale to reduce sequencing costs by several orders of magnitude.¹² The reads it produces are shorter (~100 bp) and more error-prone, but billions of them can be produced quickly. Like the long reads, the short read data can also be assembled into a new genome, albeit with more errors and gaps, owing to the difficulty of overcoming large repetitive regions with short genomic fragments.¹³ Alternatively, assuming the sequence for the reference genome and a new individual are highly similar, it is far more common (and computationally more efficient) to align the reads to the reference.¹⁴ String matching algorithms that allow mismatches, insertion, and deletions to appear in the alignments allow millions of sequenced reads to be placed on the reference genome quickly. Separate tools can then examine the alignments, looking for the presence of non-reference alleles, and using statistical approaches to call and genotype variants with high accuracy.¹⁵

Variant calling software has been successfully applied to many sequencing datasets for the discovery of DNMs. These variant callsets have provided insight into the development of drug resistance,¹⁶ the genetic architecture of some common diseases,¹⁷ and mutational rates in humans and chimpanzees with a strong paternal age effect.^{18–21} Many groups have released sophisticated software packages to facilitate these analyses and detailed instructions on their use.^{22,23}

Specificity and sensitivity are crucial metrics to consider for any variant callset. There are many approaches to establishing the specificity of a DNM callset. For select variants (typically on the order of ~100 variants from a callset), Sanger sequencing, Sequenom assays, and even targeted third-generation sequencing (i.e. PacBio sequencing, which can generate reads up to ~50,000 bp as of this writing) have been used successfully to validate mutations. Establishing sensitivity is much more difficult, and in some cases, overall sensitivity to DNMs will be much lower than one might initially appreciate. The underlying assumption that two genomes from the same species should be very similar is likely inappropriate for highly diverse populations (e.g. pathogens) or hypervariable regions (e.g. immune loci in mammals). If a haplotype present in the sample is too dissimilar to the reference, or perhaps even completely absent, read aligners may return incorrect results. Reads may align to the wrong location, the resultant mismatches mistaken for real mutations. Alternatively, they may fail to align at all, thus obscuring evidence of real variation.

1.2.2 A reference-free approach for assessing DNM sensitivity

Instead, it may be possible to establish the sensitivity of the reference-based protocol by considering how DNMs alter the genome of a sample with respect to its progenitors. Consider a site where a DNM - say, a single SNP - has occurred, as depicted in Figure 1.1. Although a single base of the genome has been altered, when the genome is divided into fixed-length words of length k , or "kmers", we find k kmers that are present in the child but absent in the parents. In this manner, DNMs can be considered generators of "novel" kmers - kmers present in the child but absent in the parents. These kmers can be used as a signal to indicate the presence of a *de novo* variant. By choosing k to be reasonably large so as to avoid analyzing short sequences that pervade the genome (half to two-thirds the length of a read will suffice), we can simply count continuous stretches of novel kmers as a proxy for the number of DNMs.

This approach gives us a powerful, reference-free mechanism to verify the results of the reference-based analysis. Sequencing of the whole genome is independent of any reference sequence that may already exist for the sample, and given sufficient coverage (Table 1.2 shows the requirements, assuming 76 bp reads and a 23 megabase genome), the raw data from a sequencing experiment should contain the full set of DNM-generated novel kmers, regardless of any mapping issues.²⁴ Taking the reads that map, calling *de novo* variants, and extracting the novel kmers from the immediate vicinity should theoretically reproduce that set. Comparing the expected (reference-free) set to the observed (reference-based) kmer set should thus provide the sought sensitivity measure.

1.3 DNM sensitivity of the reference-based approach

We now demonstrate these ideas on real data sets that will be used throughout this dissertation: experimental crosses between malaria parasites (*Plasmodium falciparum*).

1.3.1 *Plasmodium falciparum*

The datasets that we'll be referring to in this work are summarized in Table 1.3. For simplicity, we shall focus on the samples from the 3D7xHB3 cross.

1.3.1.1 Data processing

We first generated the list of expected novel kmers by processing the raw sequencing data for each sample with the *de novo* assembly software, Cortex²⁵ at $k = 47$, standard settings for other parameters, and no error cleaning. We produced the initial list of putative novel kmers by selecting kmers present in a child but absent in both parents. We further produced a "confident" list of novel kmers by imposing a kmer coverage threshold, automatically determined by a custom algorithm designed to find a local minimum on the LOESS regression of the coverage distribution, as shown

Table 1.2: Theoretical percentage of the genome recovered at a target depth of coverage.

coverage	numReads	numNucleotides	pctGenome
1	302631	2.3e+07	63.21
2	605263	4.6e+07	86.47
3	907894	6.9e+07	95.02
4	1210526	9.2e+07	98.17
5	1513157	1.15e+08	99.33
6	1815789	1.38e+08	99.75
7	2118421	1.61e+08	99.91
8	2421052	1.84e+08	99.97
9	2723684	2.07e+08	99.99
10	3026315	2.3e+08	100.00
11	3328947	2.53e+08	100.00
12	3631578	2.76e+08	100.00
13	3934210	2.99e+08	100.00
14	4236842	3.22e+08	100.00
15	4539473	3.45e+08	100.00
16	4842105	3.68e+08	100.00
17	5144736	3.91e+08	100.00
18	5447368	4.14e+08	100.00
19	5750000	4.37e+08	100.00
20	6052631	4.6e+08	100.00

Table 1.3: The *P. falciparum* datasets that will be referred to throughout this work.

	3D7xHB3	HB3xDD2	7G8xGB4	803xGB4
progeny	16	39	41	34
read length	76	76	76	100
fragment size	304 ± 32	251 ± 50	295 ± 23	222 ± 10
coverage	96 ± 40			

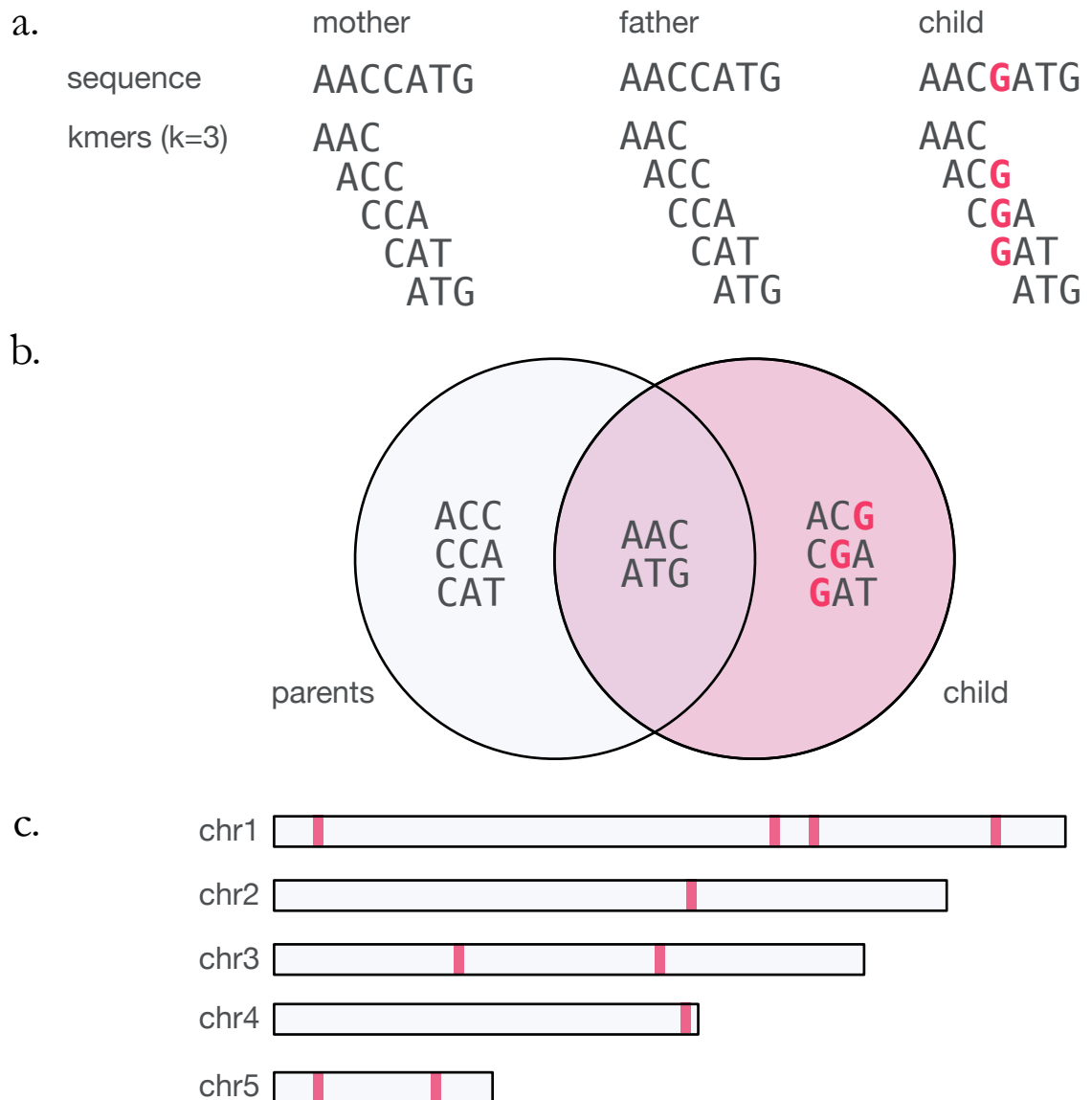


Figure 1.1: a. Parental and child sequences at the site of a *de novo* mutation, and the kmers generated at $k = 3$. b. The resulting Venn diagram of kmers found exclusively in the parents, the child, or common to both. c. Novel kmers found around the genome indicate the presence of a *de novo* mutation.

in Figure 1.2. Finally, we produce a "trusted" list of novel kmers by removing kmers originating from possible contaminants, as determined by performing a BLAST search on the confident list and removing all non-*Plasmodium* hits. These steps ensure that the list of trusted novel kmers is very restrictive.

We aligned each sample's reads to the PlasmoDB 9.0 release of the *Plasmodium falciparum* genome^{26,27} using BWA-MEM.²⁸ We followed data processing guidelines as specified in the Genome Analysis Toolkit (GATK) best-practices documentation,^{22,29} flagging duplicate reads so that they are ignored downstream, and recalibrating base quality scores using the MalariaGen data release

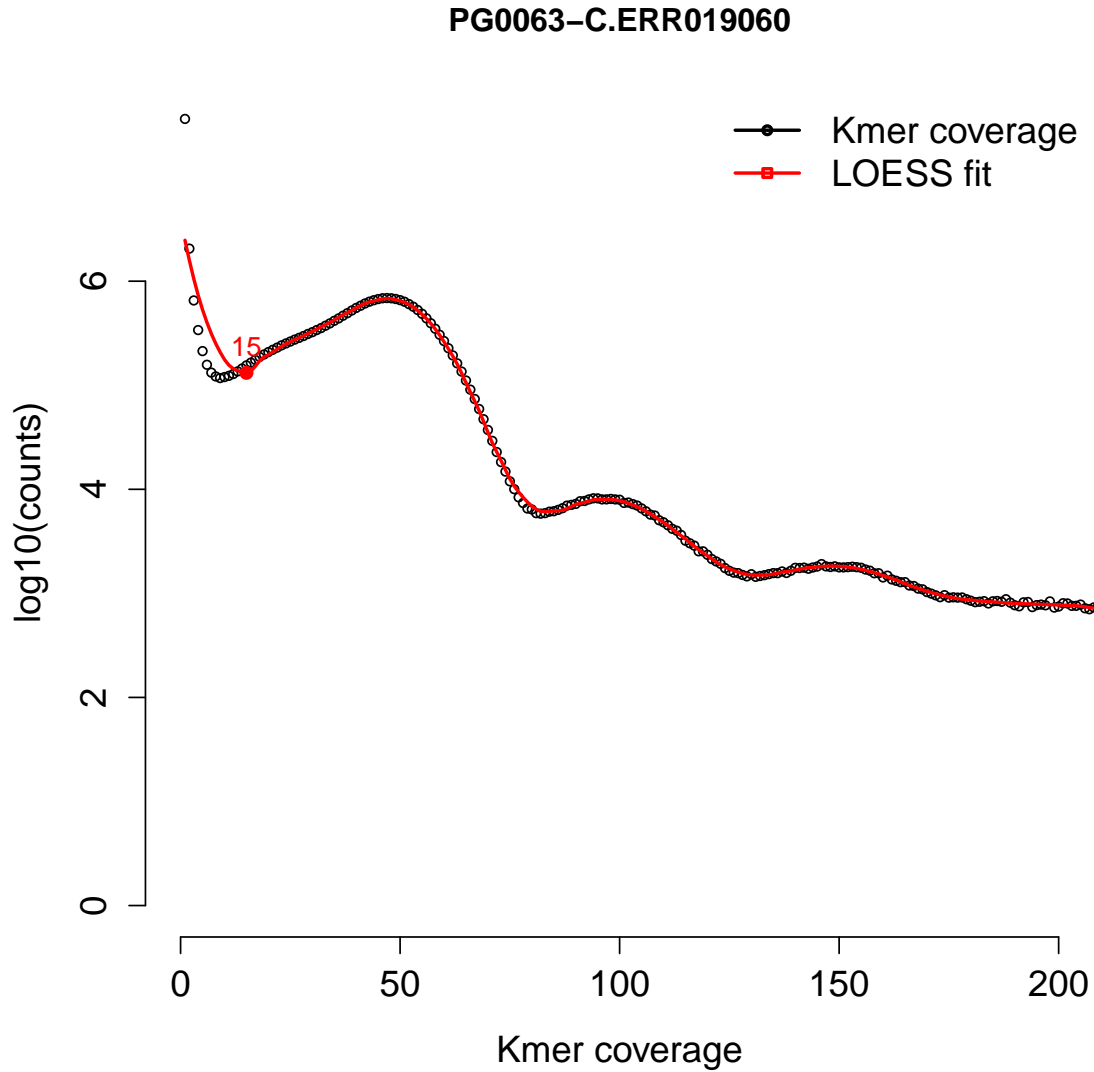


Figure 1.2: Kmer coverage distribution for a single 3D7xHB3 progeny, PG0063-C. Red line indicates non-parametric LOESS fit upon which the local minimum is detected.

on the crosses as a truth set.⁷ As recent guidance by the authors indicates the GATK's variant calling software has internalized the local indel realignment functionality, we opted to forego running this step separately. We called variants across all 18 samples (parents and progeny) simultaneously using the GATK's HaplotypeCaller with standard settings and a ploidy of 1.

A complete and perfect variant callset details the alterations that must be performed on the reference sequence in order to generate the genome of the sequenced sample. Unfortunately, it is typically not possible to obtain a perfectly sensitive and specific callset. False positives and false negatives proximal to true positives may interfere with our ability to generate the true underlying haplotype sequence. To bypass this problem, we did not filter the variant callset. Instead, we

combinatorically generate all possible subsets of variants found within 100 bp of each other using a recursive strategy to leave single elements out of a given set of kmers, presented in Algorithm 1. This will certainly generate vastly more kmers than truly exist in the sample's genome. However, as we are only interested in verifying that the variant-induced kmers are present in our trusted novel kmer set, this approach has the benefit of providing maximum sensitivity.

Algorithm 1 Given a set of variants, generate all possible subsets of variants

```

1: function GENERATEALLPOSSIBLESUBSETS(variants)
2:   loos  $\leftarrow$  []
3:   for i  $\leftarrow$  0 to length(variants) do
4:     loo  $\leftarrow$  []
5:     for j  $\leftarrow$  0 to length(variants) do
6:       if i  $\neq$  j then
7:         loo.add(variants[j])
8:       if loo.size()  $\geq$  0 then
9:         loos.add(loo)
10:      if loo.size()  $\geq$  1 then
11:        generateAllPossibleSubsets(loo)
12:   return loos

```

1.3.1.2 Results

Figure 1.3 shows the results of our reference-based vs reference-free analysis. Per sample, our restrictive reference-free analysis has generated hundreds of trusted novel kmers to explain. However, the reference-based analysis recapitulates only a fraction of these - only about 13% on average.

Attempting to explain where these missing kmers have gone, we searched the reads for every trusted novel kmer. More than 80% of reads containing these kmers were found to map to multiple homes in the genome (summarized per sample in Figure 1.4).

The preponderance of reads containing novel kmers fail to map uniquely to the reference genome, explaining why we should find such a massive discrepancy between the novel kmers we expect versus explain - no variants can be called on the unplaced reads. That there should be so many reads that fail to map is perhaps not surprising, given the divergent nature of the reference genome to other samples. Figure 1.5 shows the overlap between kmer sets between the 3D7 and HB3 genomes. More than 20% of the total set of kmers between these two samples is unique to each sample.

These unique kmers are not simply repetitive, intergenic, or otherwise uninteresting kmers. They often reflect interesting biology. Figure 1.6 shows one example: three *var* genes from 3D7's 60-member antigenic repertoire. These genes do not overlap with the HB3 repertoire. One of the 3D7xHB3 progeny, has inherited one of the 3D7 *var* genes in full, but curiously exhibits mosaic recovery of two others. This is known to be an NAHR event between the telomeres of chromosomes

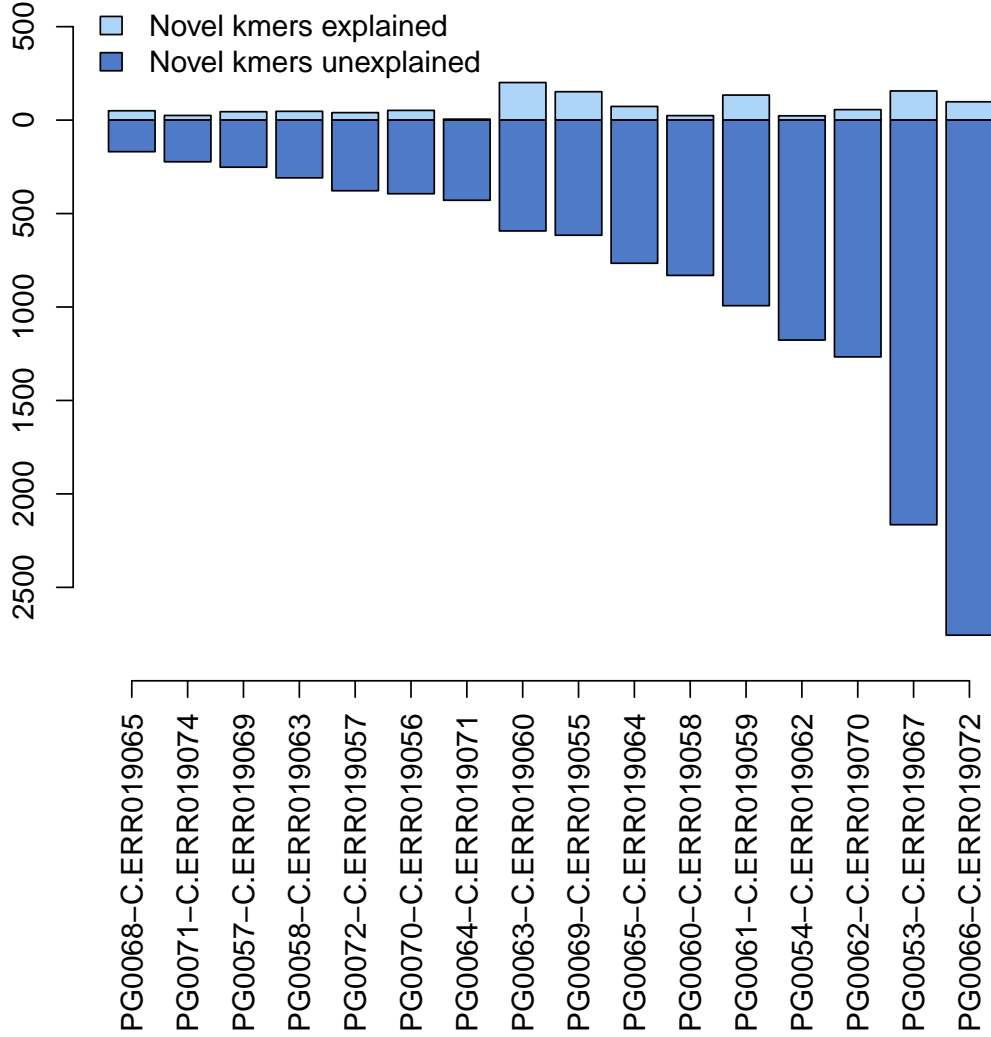


Figure 1.3: Novel kmers observed in the reference-based analysis vs trusted novel kmers expected from the reference-free analysis.

1 and 2, likely to have occurred during mitosis, preserving the domain architecture and yielding a functional product.³⁰

1.4 *De novo assembly as an alternative approach*

Alignment methods require the haplotype of a sample to closely resemble that of the reference sequence. Failing that requirement, reference-based variant callers have little recourse in recovering biologically interesting variation in genomic regions of high diversity.

Theoretically, an assembly approach will have much greater sensitivity. As it is unbiased towards the reference, we side-step the issues of the aligner failing to recognize complex genomic changes that may manifest in reads from a single locus being scattered across the genome. Fur-

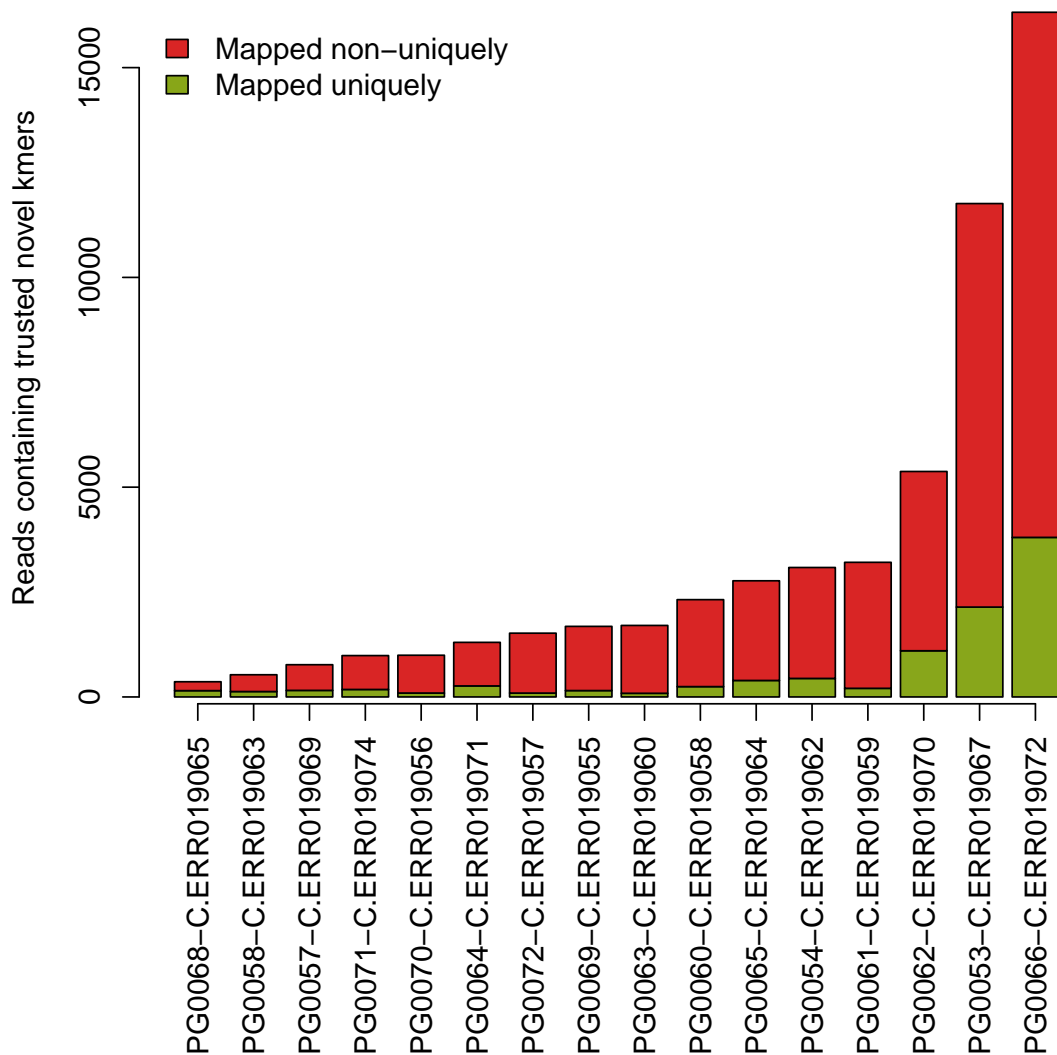


Figure 1.4: Reads that map once to the reference genome versus mapping multiple times, conditioned on the read containing a trusted novel kmer.

thermore, a single assembly algorithm should be able to discover the full gamut of point, insertion, deletion, and structural mutation present in a genome, rather than having to rely on myriad algorithms with typically poor sensitivity to such events. By restricting our processing of the assembly graphs to regions containing novel kmers, we should substantially reduce the number of potential false-positives while also providing a proxy for sensitivity, thus allowing us to determine how many regions of the graph need to be processed in order to achieve completeness of our callset.

In this dissertation, we present and demonstrate a *de novo* assembly approach to *de novo* mutation detection and genotyping. We demonstrate that this approaches yields vastly superior sensitivity and specificity, and can easily access events that occur on the haplotypic background

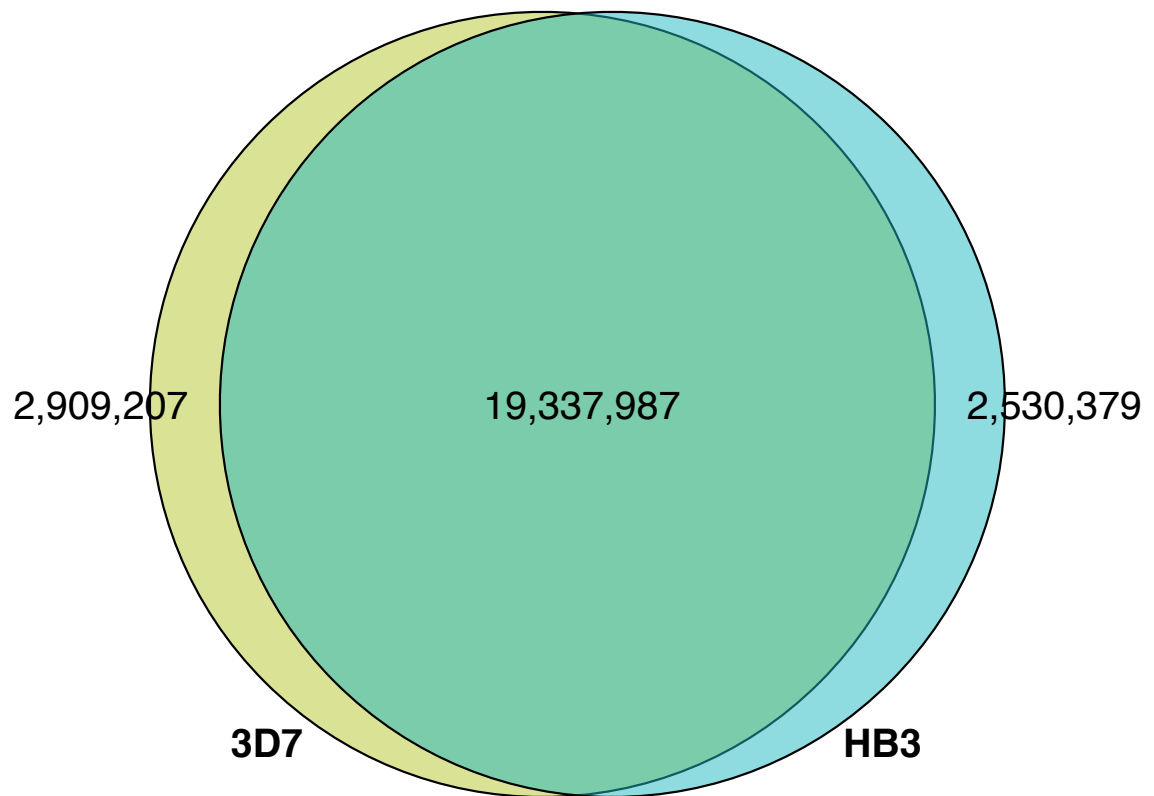


Figure 1.5: Venn diagram of kmers present in the 3D7 and HB3 genomes at $k = 47$. Both forward and reverse-complement kmers are considered the same.

of the non-reference parent.

2 *Background*

2.1 *How genome changes*

2.1.1 *Cross-over*

2.1.2 *Gene conversion*

2.1.3 *Point mutations*

2.1.4 *Structural variants*

2.1.4.1 *Small (indels)*

2.1.4.2 *Large (fusions, NAHR)*

2.1.4.3 *Chromosomal changes*

2.2 *Rates*

2.3 *Factors influencing*

2.3.1 *Replication time*

2.3.2 *Mat/pat age effects*

2.3.3 *Biases/locality*

2.4 *Known events in species*

2.4.1 *Pf.*

2.4.2 *Human*

2.4.3 *Chimp*

2.4.4 *Others*

3 *Simulation*

De novo MUTATIONS WILL UNDOUBTEDLY TAKE MYRIAD FORMS (SNPs, insertions and deletions of varying length, expansions and contractions of short tandem repeats, tandem duplications, non-allelic homologous recombinations, and possibly even inversions). Detecting all types of variants is a considerable challenge, and the software to do so will be introduced in the next chapter. In order to measure that software's expected sensitivity and specificity to such variation, we require truth datasets to which we can compare our calls. A sufficiently small genome (on the order of tens of megabases), could be run on third-generation sequencers, the long reads used to assemble full-length genomes. Then, the variants called from short-read second-generation sequencing data could be compared to the "truth" dataset established by the newer platform. However, this is still very expensive (thousands of dollars for each sample), which limits the number of samples that could be feasibly obtained. Furthermore, if variants of the classes we are attempting to identify be absent in the handful of genomes we can afford to sequence, we would not be able to accurately ascertain our power.

We chose instead to pursue a simulation strategy in order to measure our DNM variant calling performance¹. There are two components to these simulations: first we must generate the genomes of the parents and several children, including each type of DNM variant we hope to discover. From these genomes, we must then simulate reads that realistically model errors inherent in our data (matching read lengths, fragment size distributions, single base mismatch errors, indel errors, read pair chimeras, coverage profile, etc.). We discuss both of these components below.

3.1 *Simulating genomes*

Simulating a genome merely involves generating an artificial reference sequence in FASTA format. Our framework is a simple forward simulation of samples. We first generate the genomes of the parents. To generate a child's initial genome, we perform recombination *in silico*. We then add *de novo* mutations on this substrate, thus producing the child's final genome.

¹Several months after the simulation framework was completed, we obtained PacBio sequencing data on the parents of the 803xGB4 cross and three randomly selected progeny. These results will be discussed in a later chapter.

We make use of the Variant Call Format (VCF),³¹ a text file that encodes one variant per line, specifying the genomic locus, reference and alternate alleles, and metadata for the variant, to describe differences between the two parents and the DNMs to incorporate into the child’s genome. While the `FastaAlternateReferenceMaker` module in the GATK does purport to generate a new reference sequence based on variants in a VCF, we note that at the time of this writing, it silently fails to incorporate multinucleotide polymorphisms (MNPs) (simultaneous deletion and insertion). We generate many of these events to remove a reference allele and add an alternate allele in its place. To include this critical functionality, we developed our own tool to permute an existing reference sequence based on a single-sample VCF file.

Our algorithm, `IncorporateVariantsIntoGenome`, is described in Algorithm 2. Briefly, the sequence of each chromosome is loaded into an array, one nucleotide per array element. We then iterate over each record in the VCF file. For each SNP or insertion, we replace the reference nucleotide at that position with the entire alternate allele (for insertions, more than a single nucleotide). For deletions, we replace each corresponding array elements with empty strings. To generate the new genome, we iterate through each element of the array, emitting the string found in each position.

Note that we chose not to process each variant iteratively as insertions (deletions) would increase (decrease) the size of the array, altering the mapping between the VCF positions and the array positions. Keeping track of the mapping in spite of the changes is cumbersome. Instead, our scheme of placing all of the variants on the chromosome array first and then emitting the resulting sequence preserves the mapping. We will revisit this strategy later on in this chapter when we introduce an algorithm to lift data over from the reference genome coordinates to a simulated genome’s coordinates.

Finally, note that this algorithm could fail to produce a correct FASTA file if there are multiple variants that overlap. We are careful to avoid that scenario; we set our simulated variants to be placed no closer than 1000 bp from one another.

Algorithm 2 Generate an alternative reference sequence based on a VCF file.

```

1: function INCORPORATEVARIANTSINTOGENOME(ref, vcf)
2:   for all chr in ref do
3:     vcs ← vcf.getVariants(chr)
4:     for all vc in vcs do
5:       if vc.getType() == DEL || vc.getType() == MNP then
6:         for pos in vc.getPosition() : (vc.getPosition() + vc.getReferenceAllele().length()) do
7:           chr[pos] = ""
8:       chr[vc.getPosition()] = vc.getAlternateAllele()
9:   write(chr)

```

Table 3.1: Assembly statistics on publicly available finished and draft *P. falciparum* references, ordered by scaffold N50 length. Parental samples are shown in boldface.

Isolate	Origin	Length (Mb)	Scaffolds	Scaffolds N50 (Kb)	%Q40
3D7	Unknown	23.30	16	1,690.00	-
HB3	Honduras	24.26	1,189	96.47	93.17
IGH-CR14	India	21.74	849	37.02	95.49
DD2	Indochina/Laos	20.88	2,837	19.11	85.66
RAJ116	India	14.11	1,199	13.00	89.68
VS/1	Vietnam	18.89	5,856	4.42	74.79
7G8	Brazil	14.28	4,843	3.87	71.00
Senegal_V34.04	Senegal	13.24	4,329	3.76	76.22
D10	PNG	13.38	4,471	3.71	71.80
RO-33	Ghana	13.71	4,991	3.47	69.91
K1	Thailand	13.29	4,772	3.42	73.30
FCC-2/Hainan	China	12.96	4,956	3.30	69.39
D6	Sierra Leone	13.22	5,011	3.23	71.62
SL	El Salvador	13.19	5,193	3.08	69.41
PFCLIN	Ghana	42.19	18,711	2.99	-

3.1.1 Parents

We began by simulating the genomes of two parents using a workflow depicted in Figure 3.1. For convenience, we simulated samples from the 3D7xHB3 cross. Choosing the existing reference sequence for the 3D7 sample obviates the need to generate any data for the first parent. The second parent is trickier; while an existing draft reference sequences does exist for HB3, it is of low quality, assembled into thousands of pieces rather than the simple 14 autosomes we expect (Table 3.1 shows metrics on every *P. falciparum* sample publicly available). Using the supercontigs from draft reference sequences is hugely cumbersome for simulating recombination as it is not straightforward to decide which chromosomes in the 3D7 genome and which supercontigs should be processed together.

Instead, we chose to produce a new HB3 reference genome sequence by taking the 3D7 reference and inserting the appropriate modifications using Algorithm 2. These modifications are comprised of two parts: introducing the appropriate variants, and replacing the *var* gene repertoire.

We first obtained a VCF of variants found in the HB3 sample, PG0052-C, from the Malaria-Gen 3D7xHB3 cross dataset.³² Variant counts are described in the Table 3.2, and include SNPs, insertions, deletions, and complex (simultaneous insertions and deletions) events. These calls were made by combining the reference-based UnifiedGenotyper module in the GATK²² and the reference-free bubble-calling algorithm in the Cortex²⁵ software. All calls were restricted to the core genome; subtelomeric regions were masked out due to poor mapping properties (owing to the tremendous diversity in these regions of other *P. falciparum* parasites with respect to the 3D7 reference).

a. Call variants in one parent (HB3) against the other (3D7)



b. Delete 3D7 *var* genes, insert compatible HB3 counterparts in their place



c. Replace reference alleles in 3D7 with variant alleles



Figure 3.1: Workflow for generating the HB3 parental genome. a. Reads from HB3 sample, PG0052-C, are mapped to the 3D7 reference genome, and variants (SNPs and indels) are called and stored as a VCF file. b. We remove the 3D7 *var* gene repertoire, replacing each with a reasonable HB3 *var* counterpart, and encode the changes to the 3D7 reference genome as a VCF file. c. We alter the reference genome using Algorithm 2, thus producing the simulated HB3 genome.

Table 3.2: Variants found the HB3 (PG0052-C) sample from the MalariaGen 3D7xHB3 dataset.

variants	SNPs	insertions	deletions	complex
42,054	15,376	11,807	14,643	228

Next, we produced a VCF describing *var* gene replacements. The sequences for HB3 *var* genes and upstream promoter metadata were obtained from the VarDom server.³³ No positional information from this data source is available, thus the exact placement of these *var* genes in a chromosomal context is unclear. However, previous work has established a strong association between conserved sequences of upstream promoters (phylogenetically grouped into five classes: A through E) and placement in the genome.³⁴ We therefore replaced 3D7 *var* genes with HB3 *var* gene counterparts, taking care to replace genes with similar UPS classes whenever possible, and grouping genes with suspiciously incomplete metadata otherwise. The replacements were described in the resulting VCF as simultaneous deletions of the 3D7 allele and insertions of the HB3 allele. No effort was made to match the orientation of the replacement *var* gene with the replaced *var* gene. The precise replacements are summarized in Table 3.3.

These two VCFs were combined to produce a complete set of changes required to transform the 3D7 genome into a pseudo HB3 genome. The transformation was applied using Algorithm 2.

Table 3.3: *Var* gene replacements from 3D7 to HB3 repertoire.

3D7 gene	3D7 ups class	HB3 gene	HB3 ups class
PF3D7_0421100	UPSB5	PFHG_02500	UNKNOWN
PF3D7_0600200	UPSB2	PFHG_02495	UNKNOWN
PF3D7_0632500	UPSB5	PFHG_05132	UNKNOWN
PF3D7_0800300	UPSB2	PFHG_04012	ND
PF3D7_1200400	UPSB5	PFHG_05200	ND
PF3D7_1240900	U	PFHG_05483	ND
PF3D7_0400400	UPSA1	PFHG_03840	UPSA1
PF3D7_0425800	UPSA1	PFHG_03671	UPSA1
PF3D7_1100200	UPSA1	PFHG_04861	UPSA1
PF3D7_1150400	UPSA1	PFHG_05052	UPSA1
PF3D7_1300300	UPSA1	PFHG_03234	UPSA1
PF3D7_0533100	UPSA2	PFHG_03521	UPSA2*
PF3D7_0100300	UPSA3	PFHG_02274	UPSA3
PF3D7_0100100	UPSB1	PFHG_04081	UPSB1
PF3D7_0115700	UPSB1	PFHG_04749	UPSB1
PF3D7_0200100	UPSB1	PFHG_03516	UPSB1
PF3D7_0223500	UPSB1	PFHG_04277	UPSB1
PF3D7_0300100	UPSB1	PFHG_03717	UPSB1
PF3D7_0324900	UPSB1	PFHG_04035	UPSB1
PF3D7_0400100	UPSB1	PFHG_04491	UPSB1
PF3D7_0426000	UPSB1	PFHG_04620	UPSB1
PF3D7_0500100	UPSB1	PFHG_04593	UPSB1
PF3D7_0632800	UPSB1	PFHG_04770	UPSB1
PF3D7_0700100	UPSB1	PFHG_04057	UPSB1
PF3D7_0712300	UPSB1	PFHG_03232	UPSB1
PF3D7_0733000	UPSB1	PFHG_04928	UPSB1
PF3D7_0800100	UPSB1	PFHG_03416	UPSB1
PF3D7_0413100	UPSB3	PFHG_03476	UPSB3*
PF3D7_0712400	UPSB3	PFHG_02421	UPSB3
PF3D7_1240300	UPSB4	PFHG_02272	UPSB4
PF3D7_0809100	UPSB6	PFHG_02276	UPSB6
PF3D7_0712800	UPSB7	PFHG_04014	UPSB7
PF3D7_0808700	UPSB7	PFHG_02425	UPSB7
PF3D7_1240400	UPSB7	PFHG_04769	UPSB7
PF3D7_0412400	UPSC1	PFHG_03480	UPSC1
PF3D7_0412700	UPSC1	PFHG_03478	UPSC1
PF3D7_0412900	UPSC1	PFHG_00592	UPSC1
PF3D7_0420700	UPSC1	PFHG_02419	UPSC1
PF3D7_0420900	UPSC1	PFHG_02429	UPSC1
PF3D7_0421300	UPSC1	PFHG_02277	UPSC1
PF3D7_0617400	UPSC1	PFHG_02273	UPSC1
PF3D7_0712900	UPSC2	PFHG_04015	UPSC2
PF3D7_1200600	UPSE	PFHG_05046	UPSE

3.2 Children

Generating the genomes of the children is slightly more involved, as there is much more biology to simulate, and many more considerations to be made when placing variants. We generated VCF descriptions of the children using a multi-stage pipeline that simulates (in order):

1. homologous recombination between 3D7 and HB3
2. gene conversions
3. NAHR events between *var* genes
4. *de novo* SNPs, insertions, deletions, and inversions

3.2.1 Homologous recombination

3.2.1.1 Allelic

Recombination frequency will be dependent on the length of the chromosome. The empirical distributions were generated from the 75 samples in the crosses data. These are plotted in Figure 3.2. We simulated the number of recombination events per chromosome by drawing a random number from these distributions. The recombination sites themselves were chosen by drawing a uniform random variate between 1 and the length of the chromosome. Although there are certainly hotspots and coldspots of recombination in the genome, we have ignored this complication. An example haplotype mosaics for chromosome 12 is found in Figure 3.3.

3.2.1.2 Gene conversion

To simulate gene conversion events, we first chose a handful of random sites known to be variant in HB3, and determined the size of the event (number of adjacent HB3 variants involved in the gene conversion) by choosing a random uniform variate between 1 and 3. If these sites were originally transmitted to the child, they were removed from the VCF. If they had not been transmitted, they were added. The homologous recombinations and gene conversion events are displayed below for each chromosome and sample.

3.2.1.3 Non-allelic

NAHR events were generated by first finding compatible recombination partners. This list was generated by determining which 3D7 *var* genes had been transmitted to the child, grouped by ups class and telomeric positioning (5' or 3'). In each group, two random genes were chosen. If necessary, the gene sequences were reverse complemented to have matching orientation. As NAHR events between two *var* genes appear to occur in regions of homology, we identified shared 9-mers, positioned between 20 bp and 100 bp between the two genes, to act as possible recombination sites.

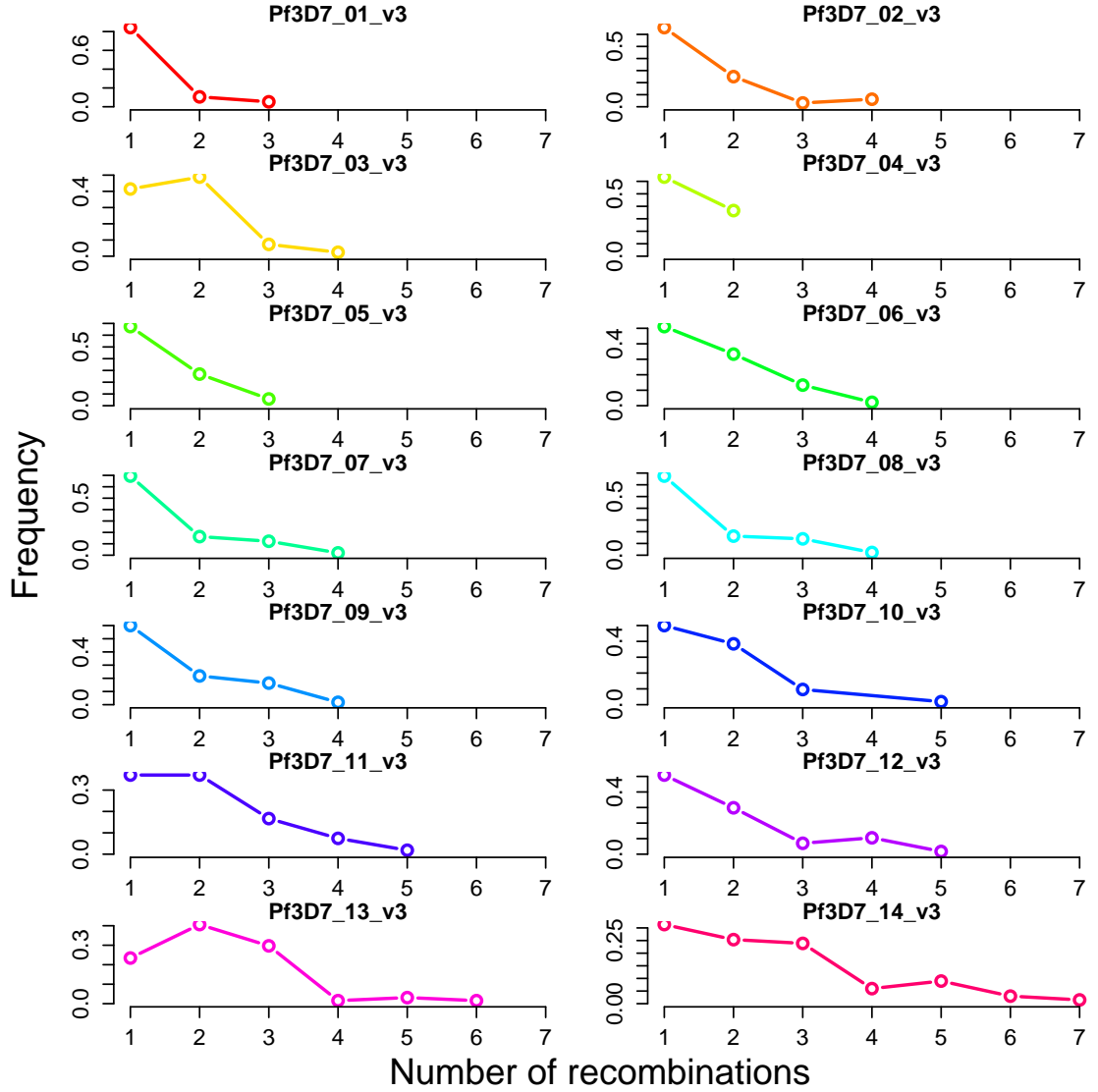


Figure 3.2: Empirical recombination frequencies per chromosome

We selected between 2 and 5 random switches between the sequences and copied sequence data accordingly. Keeping in mind that our previous work has shown that the recombined *var* genes are lost in order to produce the chimera, we added VCF records to delete the previous *var* alleles from the child's genome and replace one of them with the recombined sequence.

3.2.2 SNPs and indels

To simulate de novo SNPs, we added 250 sites with random (non-reference) alleles at random positions throughout the genome. We also simulated 10 insertion, deletion, and inversion events at every length between 1 and 100 bp. For insertions, random alleles were generated and tested to ensure they did not match the reference sequence. For deletions, we simply replaced the reference

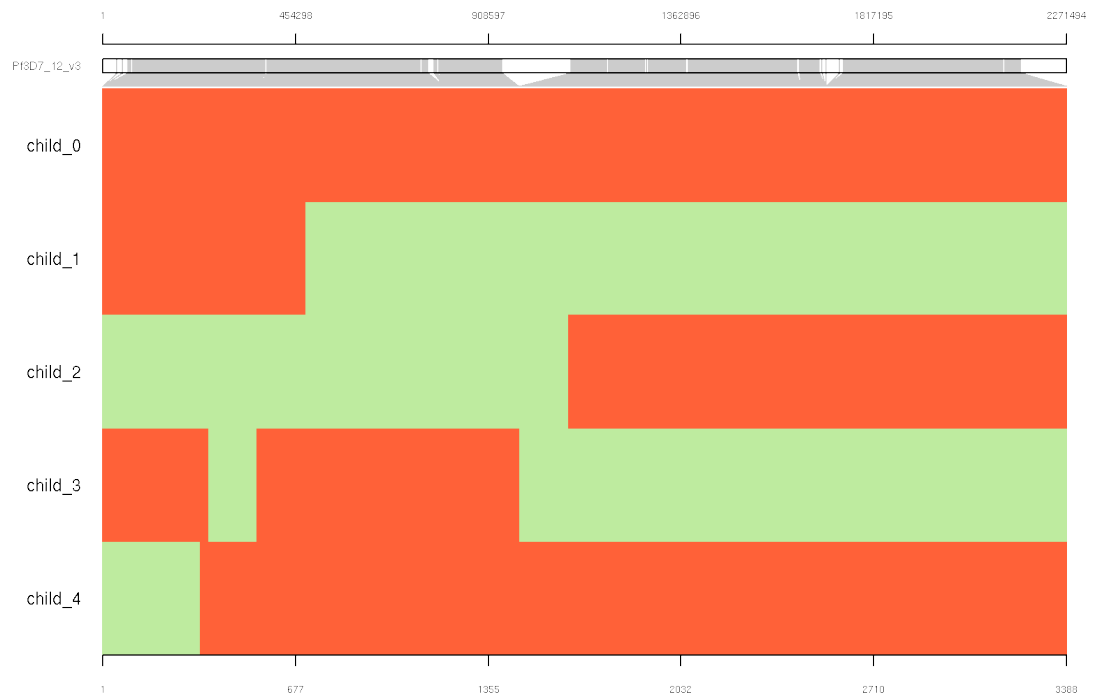


Figure 3.3: Simulated haplotype mosaics for chromosome 12.

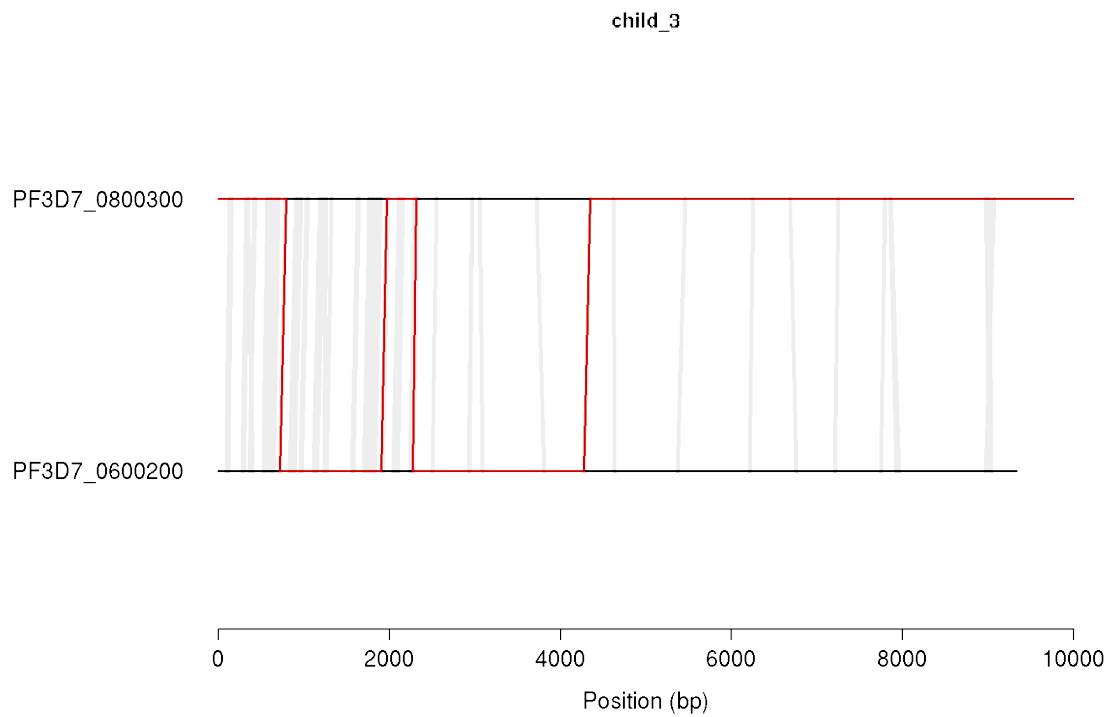


Figure 3.4: Non-allelic recombinations for two compatible *var* genes.

allele with a truncated allele of the prescribed length. For inversions, we replace the reference allele with its complement.

3.2.2.1 Expansion of short tandem repeats

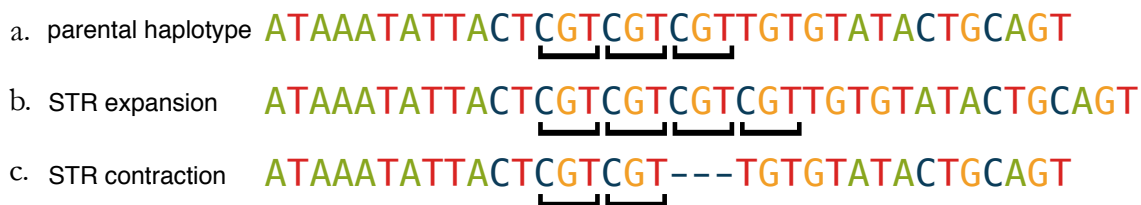


Figure 3.5: Expansion and contraction of short tandem repeats (STRs)

As a special case of indels, we sought specifically to simulate the expansion and contraction of short tandem repeats (STRs), depicted in Figure 3.5. As these are constrained to occur at loci where there are already existing repeats, we first built a map of repeated 2-bp, 3-bp, 4-bp, and 5-bp sequences in the 3D7 genome. We filtered these lists, retaining only STRs where the repeated unit occurred at least three times. For each simulated event, we randomly chose a position from the appropriate list and select a number of units to add or remove. The number of units is constrained to be less than the length of the number of repeat units of the existing STR. With these considerations, we simulated 10 expansions and 10 contractions at the aforementioned repeat unit sizes.

3.2.2.2 Tandem duplications

For tandem duplications, we simulated 10 events at each length between 10 bp and 50 bp by selecting sequences randomly from the genome and inserted an identical copy at the same locus.

3.3 Simulating reads

We now turn our attention to simulating second-generation sequencing reads given an underlying genome. There are existing tools that will simulate perfect reads and uniform coverage, which will be important for initial tests of our variant identification software. However, the crucial simulation is of imperfect reads with non-uniform coverage. Without this, any estimate of our sensitivity and specificity is unlikely to be predictive of performance in real data.

There are many tools that can simulate imperfect reads from a given sequence. Almost every solution involves user-specified parameters controlling the properties of the sequencing data. For instance, a tool may model fragment size as a normal distribution, requiring the user to specify the requisite shape parameters. It may also permit the user to specify a desired mismatch rate to simulate the presence of sequencing error. Some tools have presets that automatically set parameters to those consistent with average behavior for various sequencing platforms.

The problem with all of these tools is an over-reliance on assumed parameters of real data. Should a fragment size distribution for real data deviate from the typical normal distribution, this

will not be captured in the simulated dataset. Mismatch and indel errors do not happen at any position in the read with equal probability, but rather are biased towards later cycles and certain sequence contexts. Read coverage is not simply Poisson-distributed, but varies across the length of the genome depending on GC bias, secondary structure, even the particular sequencing chemistry used. There are currently no tools that are capable of capturing such nuance.

Instead, we chose to learn empirical distributions of major sequencing properties using an exemplar dataset and sample from those distributions directly in order to generate reads. This frees us from having to make any assumptions about our dataset, easily generalizes to any dataset regardless of when, where, or how it was sequenced. It's also vastly more realistic than other approaches.

Our approach is detailed below. Briefly, it consists of three components:

1. A "coverage profile": a description of where every fragment and read in the genome should fall and which should contain errors.
2. A "read profile": a description of where to place errors within a read, including mismatches, insertions, and deletions.
3. The read simulator: samples from the profiles to generate reads in the new reference sequence.

3.3.1 Constructing the coverage profile

Construction of a coverage profile consists of two sub-problems: providing a complete description of where reads fall on the existing reference sequence, and transferring that information sensibly to the modified reference sequence. We address each of these needs with custom tools.

3.3.1.1 ComputeBaseAndFragmentErrorRates

The `ComputeBaseAndFragmentErrorRates` tool provides a precise specification for the number of fragments and reads found starting at every position in the canonical reference genome, as well as an accounting as to which reads and fragments contained any kind of error (mismatch or indels). Briefly, we iterate over every chromosome in the reference genome, advancing through every aligned read stored in an exemplar sample's BAM file. We instantiate a chromosome-length array indicating the number of reads that start at a given position and the number of reads that contain apparent errors (any discrepancy from the reference sequence).

We must also store information about read fragment starts and error rates, requiring us to keep track of read pairs as we traverse the BAM file. To do so, we hash the read name to the first instance of the read we see along the length of a chromosome. As paired-end reads have the same name, the second time we see the same read name, we will have found the second end of the pair.

We then increment the count at the chromosome array position that corresponds to the 5'-most end of the pair, and increment the fragment errors array based on errors in either end of the pair.

This algorithm is described in Algorithm 3.

Algorithm 3 Emit all fragment starts, read starts, and error rates per position.

```

1: function COMPUTEBASEANDFRAGMENTERRORRATES(BAM)
2:   for chr in chrs do
3:     numReadsWithErrors  $\leftarrow$  []
4:     numReads  $\leftarrow$  []
5:     numFragmentsWithErrors  $\leftarrow$  []
6:     numFragments  $\leftarrow$  []
7:     seenReads  $\leftarrow$  []
8:     reads  $\leftarrow$  BAM.getAllReads(chr)
9:     for read in reads do
10:      readErrorPositions  $\leftarrow$  getPositionsWithErrors(read)
11:      refErrorPositions  $\leftarrow$  convertReadPositionsToReferencePositions(readErrorPositions)
12:      for refErrorPosition in refErrorPositions do
13:        numReadsWithErrors[refErrorPosition] ++
14:      for readPosition in 1 : read.length() do
15:        numReads[convertReadPositionsToReferencePositions(readPosition)] ++
16:      if !seenReads.contains(read.getName()) then
17:        seenReads  $\leftarrow$  read.getName()
18:      else
19:        mateErrorPositions  $\leftarrow$  getPositionsWithErrors(seenReads[read.getName()])
20:        if then readErrorPositions.size() + mateErrorPositions.size()  $\geq$  1
21:          numFragmentsWithErrors[seenReads[read.getName()].getAlignmentStart()] ++
22:          numFragments[seenReads[read.getName()]] ++
23:      for pos in 1 : numReads.length() do
24:        print chr, pos, numReadsWithErrors[pos], numReads[pos], numFragmentsWithErrors[pos], numFragments[

```

3.3.1.2 LiftoverFromRefToChild

The genomic coordinates present in the table produced by Algorithm 3 must be transformed from the reference sequence to the simulated genome before it can be used. To do so, we developed a tool that could liftover the coordinates appropriately when given the table, reference sequence, and a VCF file describing the alterations made to the reference to transform it into the simulated genome. This is accomplished as follows: we first iterate through each chromosome and store the entire sequence as an array of strings. Each string is of variable length, and at this stage of the construction, all stored strings are of length 1: the length of the reference allele present at that location. At positions where a variant is present, we make alterations to the array. We first replace reference nucleotide strings with zero-length strings up to the length of the reference allele. We then set the string at the variant position to the alternate allele indicated in the VCF file. Thus, we are able to handle SNPs (one reference base blanked out, then replaced with one alternate base), insertions (one base blanked out, many replaced), deletions (many bases blanked

out, one replaced), and MNPs (many bases blanked out, many replaced). When all of the strings are concatenated together, the new chromosome is constructed.

Once the chromosome is generated, we advance through the sequence one position at a time. At each position, we emit the contents of the previously constructed table with the number of reads, fragments, and errors for each. At some positions, the presence of variants will have modified the underlying sequence such that new positions do not have explicit read and fragment information to lift over. To account for this, we keep running statistics on previously seen read and fragment statistics from which we can compute running means and standard deviation. We generate new values for the number of reads, fragments, and error rates as necessary as the mean plus standard deviation of the relevant metric, ensuring the values are never less than 0, and that we round up or down to the nearest integer (error rates are rounded up and read/fragment counts are rounded down to increase our self-penalty).

This algorithm is described in ??.

3.3.2 *Constructing the read profile*

Next, we generate the read profile, describing the various properties of fragments and read. As in previous examples, we iterate over each read in the exemplar sample's BAM file, storing relevant information in tabular form.

3.3.2.1 *Scalar properties*

We first store the following scalar information (as the data we are modelling is assumed to be Illumina data generated from a single library, some properties which could otherwise be stored as distributions are instead treated as a single number that will be simple constants in our simulation):

1. read length
2. number of reads
3. number of reads with errors
4. number of read pairs
5. number of chimeric pairs (paired-end reads with one end aligned to a different chromosome than the other)
6. number of mismatches
7. number of insertions
8. number of deletions

Note that some of the mismatches, insertions, and deletions may represent true variation between the sample and the reference sequence. We do not mask these sites out. While this increases the apparent error rate, the variant rates are typically low compared to the number of bases in the reference sequence itself, making the difference negligible. Furthermore, we will choose the reference sample as the exemplar dataset for the read simulations. Since this sample should theoretically be identical to the reference sequence, this choice obviates the need for any masking of variants in the reference sample against the reference sequence.

3.3.2.2 *Empirical distributions*

Other properties take on a range of values with some probability. Rather than trying to fit the underlying distributions explicitly (which can often fail as datasets contain more nuance than what sequencing platforms should theoretically produce), we instead store empirical distributions and generate random values from them accordingly. We store the following empirical distributions:

1. number of errors (of any type - mismatch, insertion, or deletion) in a read
2. fragment size
3. insertion size
4. deletion size

Note that these distributions are not conditioned on position in the read.

3.3.2.3 *Covariate table*

Finally, we construct the covariate table: a set of empirical distributions for various error events conditioned on the following:

1. type (SNP, insertion, deletion)
2. end of pair (first end or second end)
3. strand (positive or negative)
4. 5' dinucleotide context
5. position in read
6. first base of error sequence (empty for deletions)

For each read, we increment an element in a table that corresponds to these six covariates. The code to do so is contained in our `GenerateReadSimProfile` module.

3.3.3 *The read simulator*

With the coverage and read profiles in hand, we are finally ready to simulate reads. We advance through each base of the simulated genome, reading the coverage profile as we traverse. At each site, we use the coverage profile to dictate the number of fragments that must be generated and how many of those fragments should contain errors. We then

4 *Detection and genotyping*

4.1 *Introduction*

In the preceding chapters, we discussed why alignment-based methods may fail to detect *de novo* mutations in real datasets, the types of variation we might encounter, tools for simulating genomes with these events, and tools for simulating idealized and realistic NGS data from these altered genomes. We now turn our attention to a new graph-based method for detecting and genotyping such events. The simulation framework we described will be used to establish the method's sensitivity and specificity.

4.2 *Variant motifs*

Just as reference-based methods will search for motifs in the data representing variants (e.g. mismatches, gaps, or unusual truncations in the read alignments; read pairs aligning much further apart than expected; chimeras or inter-chromosomal alignments; etc.), so must we scan for indicative motifs in the assembly graph. Before we discuss the precise nature of these motifs, it is useful to draw a distinction between "simple" and "complex" variants. A "simple" variant is a SNP, insertion, or deletion that occurs within a single chromosome. A "complex" variant is a homologous or non-homologous recombination, translocation, or other interchromosomal exchange. The patterns inherent to these two categories of variants are quite distinct.

4.2.1 *Simple variant motifs*

Detection of simple variants in *de novo* assembly data is typically described as identifying so-called "bubbles" in the de Bruijn graph: regions where a variant has broken the homology between sequences, resulting in flanking kmers that are shared between the samples, and spanning kmers that differ through the variant itself. In a single diploid sample, this could be a heterozygous SNP or indel between two homologous chromosomes. In haploid samples, one or more samples may differ from the others, resulting in the bubble.

As a simple illustration, consider three sequences from a mother-father-child pedigree, shown in figure 4.1a. While the maternal and paternal haplotypes (green and blue, respectively) are identical, the child's (red) differs by a single C to G SNP. Figure 4.1b shows the resulting multi-color de

Bruijn $k = 3$ graph built from this data. The mutation has given rise to the canonical bubble motif in the graph. Three novel kmers (kmers present in the child and absent in the parents) spanning the variant allele are present. Figure 4.2 is an equivalent graph for another simulated SNP, shown with more context and constructed with a much larger value of k appropriate for 76 - 100 bp read lengths, typical of NGS datasets (in this case, $k = 47$).

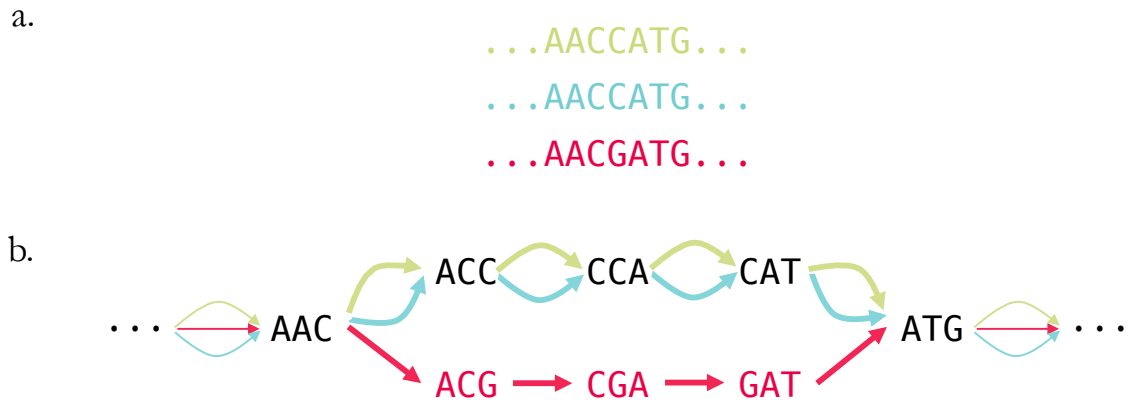


Figure 4.1: a. Haploid sequences from a mother (green), father (blue), and child (red), the last differing from the first two by a single SNP. b. The resulting multi-color de Bruijn graph for $k = 3$. Red vertices denote kmers that are deemed "novel", i.e. present in the child and absent in the parents. Edge colors reflect the samples in which the connected pairs of kmers are found. Edges that are part of the bubble (variant call) are displayed with thicker lines.

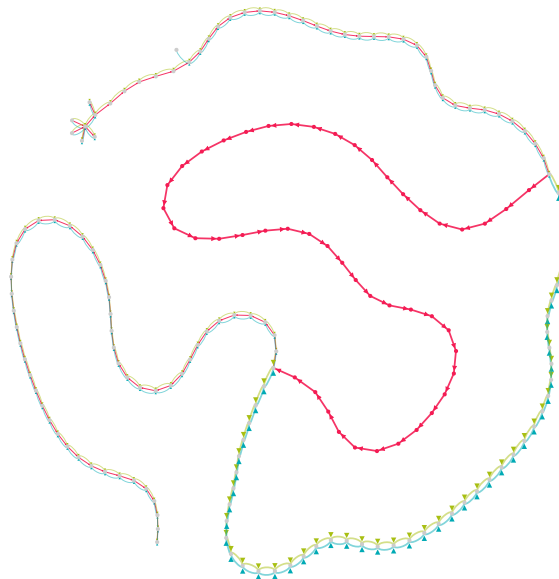


Figure 4.2: A multi-color de Bruijn graph at $k = 47$ for a haploid pedigree spanning a simulated *de novo* SNP. Vertex labels have been suppressed for clarity. Spatial layout is arbitrary and for display purposes only.

All simple variants will have this basic structure: a bubble in the graph that separates the variant samples from the non-variant samples. The only major difference is the length of each branch: longer for an insertion in the child, shorter for a deletion (note that for short events, this

is generally not apparent from the display, as evidenced by figures 4.3 and 4.4).

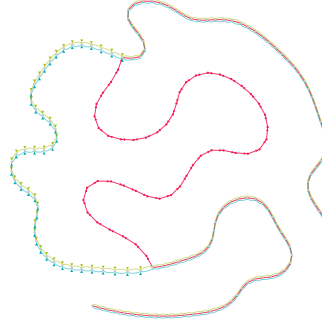


Figure 4.3: A 5 bp insertion in the child

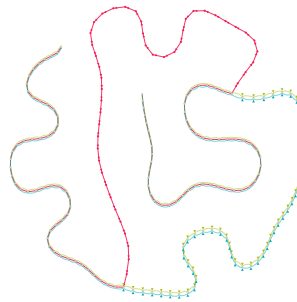


Figure 4.4: A 5 bp deletion in the child

The previous examples have all involved *de novo* variants on perfectly homologous parental haplotypic backgrounds. However, many variants may occur on the background of one parent or another. Figure 4.5 depicts one such event. A 41-bp tandem duplication has occurred on the background of the mother, as evidenced by the presence of green edges, but no blue edges. In the flanking tails, edges shared between all three samples are present, until a blue edge separates from the graph and connects to different vertices. While not shown, these branches continue along the genome of the father.

Finally, it is possible to encounter variants where the path through the graph taken by the child can appear to follow both the variant and non-variant paths, as demonstrated by figure 4.6. Such a scenario may arise by a mutation on a sequence with copy number greater than 1: both the unaltered and altered sequences would then exist simultaneously in the child's genome.

4.2.2 *Complex variant motifs*

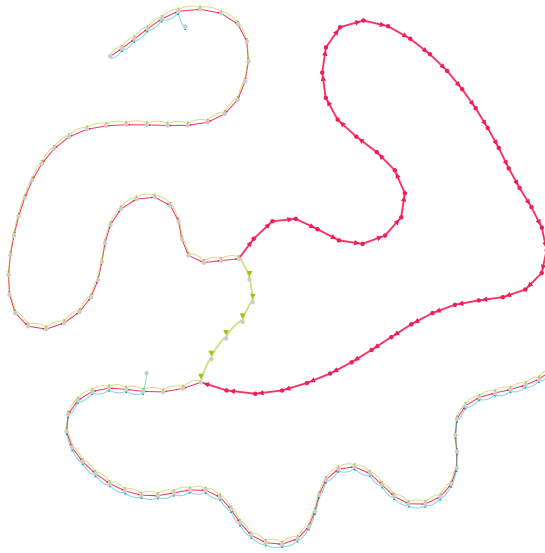


Figure 4.5: A tandem duplication on the haplotypic background of the mother.

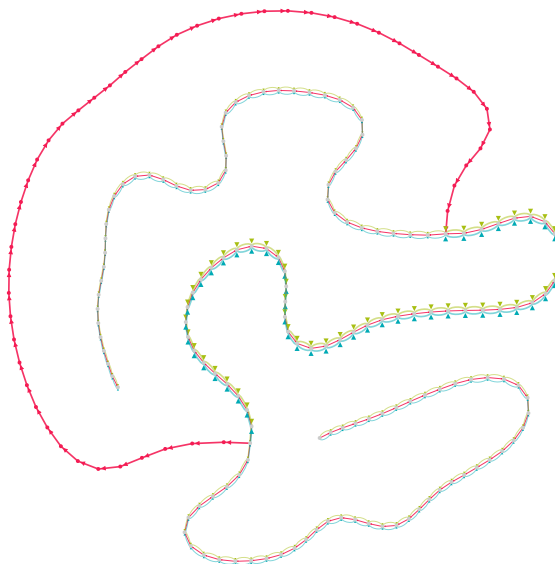


Figure 4.6: A variant wherein the child's path does not simply diverge from that of the parents, but rather navigates both.

5 *Pf*

5.1 *Lit review*

5.1.1 *Review of Kong et al., 2002*

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on 1,257. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

6 *Chimp*

6.1 *Lit review*

6.1.1 *Review of Kong et al., 2002*

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on 1,257. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

7 *Discussion*

7.1 *Lit review*

7.1.1 *Review of Kong et al., 2002*

Augustine Kong et al. discuss a new genetic map of recombination rates using genotyping information from 869 individuals in 146 Icelandic families. This is the first such map made after the sequencing of the human genome, and is thus able to leverage the new reference sequence in order to correctly order the genotyped markers. It is a substantially higher-resolution map than provided by the former gold-standard, the Marshfield map. The Marshfield map contained data on only 188 meioses, whereas the Kong et al. map contained data on 1,257. The new map reveals marked differences in recombination rates between males and females (e.g. the recombination rate in female autosomes is a factor of 1.65 higher than that observed in males) for reasons beyond sequence features.

References

- [1] Thomas S Rask, Daniel A Hansen, Thor G Theander, Anders Gorm Pedersen, and Thomas Lavstsen. Plasmodium falciparum Erythrocyte Membrane Protein 1 Diversity in Seven Genomes – Divide and Conquer. *PLoS computational biology*, 6(9):e1000933, September 2010.
- [2] William Bateson and Miss E R Saunders. Experimental Studies in the Physiology of Heredity, 1908.
- [3] T H Morgan. Sex Limited Inheritance in Drosophila. *Science (New York, NY)*, 32(812):120–122, July 1910.
- [4] D Walliker, I Quakyi, T Wellems, T McCutchan, A Szarfman, W London, L Corcoran, T Burkot, and R Carter. Genetic analysis of the human malaria parasite Plasmodium falciparum. *Science (New York, NY)*, 236(4809):1661–1666, June 1987.
- [5] D S Peterson, D Walliker, and T E Wellems. Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proceedings of the National Academy of Sciences of the United States of America*, 85(23):9114–9118, December 1988.
- [6] T E Wellems, L J Panton, I Y Gluzman, V E do Rosario, R W Gwadz, A Walker-Jonah, and D J Krogstad. Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross. *Nature*, 345(6272):253–255, May 1990.
- [7] A B Vaidya, O Muratova, F Guinet, D Keister, T E Wellems, and D C Kaslow. A genetic locus on Plasmodium falciparum chromosome 12 linked to a defect in mosquito-infectivity and male gametogenesis. *Molecular and biochemical parasitology*, 69(1):65–71, January 1995.
- [8] T Furuya, J Mu, K Hayton, A Liu, J Duan, L Nkrumah, D A Joy, D A Fidock, H Fujioka, A B Vaidya, T E Wellems, and X z Su. Disruption of a Plasmodium falciparum gene linked to male sexual development causes early arrest in gametocytogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16813–16818, November 2005.

- [9] L H Freitas-Junior, E Bottius, L A Pirrit, K W Deitsch, C Scheidig, F Guinet, U Nehrbass, T E Wellems, and A Scherf. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*, 407(6807):1018–1022, October 2000.
- [10] Michael F Duffy, Timothy J Byrne, Celine Carret, Alasdair Ivens, and Graham V Brown. Ectopic recombination of a malaria var gene during mitosis associated with an altered var switch rate. *Journal of molecular biology*, 389(3):453–469, June 2009.
- [11] E W Myers. Toward simplifying and accurately formulating fragment assembly. *Journal of computational biology : a journal of computational molecular cell biology*, 2(2):275–290, 1995.
- [12] Elaine R Mardis. A decade’s perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011.
- [13] M C Schatz, A L Delcher, and S L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173, September 2010.
- [14] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(11s):S6–S12, November 2009.
- [15] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, 12(6):443–451, June 2011.
- [16] N Woodford and M J Ellington. The emergence of antibiotic resistance by mutation. *Clinical Microbiology and Infection*, 13(1):5–18, 2007.
- [17] Benjamin M Neale, Yan Kou, Li Liu, Avi Ma’ayan, Kaitlin E Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, Paz Polak, Seungtae Yoon, Jared Maguire, Emily L Crawford, Nicholas G Campbell, Evan T Geller, Otto Valladares, Chad Schafer, Han Liu, Tuo Zhao, Guiqing Cai, Jayon Lihm, Ruth Dannenfelser, Omar Jabado, Zuleyma Peralta, Uma Nagaswamy, Donna Muzny, Jeffrey G Reid, Irene Newsham, Yuanqing Wu, Lora Lewis, Yi Han, Benjamin F Voight, Elaine Lim, Elizabeth Rossin, Andrew Kirby, Jason Flannick, Menachem Fromer, Khalid Shakir, Tim Fennell, Kiran Garimella, Eric Banks, Ryan Poplin, Stacey Gabriel, Mark DePristo, Jack R Wimbish, Braden E Boone, Shawn E Levy, Catalina Betancur, Shamil Sunyaev, Eric Boerwinkle, Joseph D Buxbaum, Edwin H Cook, Bernie Devlin, Richard A Gibbs, Kathryn Roeder, Gerard D Schellenberg, James S Sutcliffe, and Mark J Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, May 2012.
- [18] Donald F Conrad, Jonathan E M Keebler, Mark A DePristo, Sarah J Lindsay, Yujun Zhang, Ferran Casals, Youssef Idaghdour, Chris L Hartl, Carlos Torroja, Kiran V Garimella, Martine

- Zilversmit, Reed Cartwright, Guy A Rouleau, Mark Daly, Eric A Stone, Matthew E Hurles, Philip Awadalla, and 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7):712–714, July 2011.
- [19] Oliver Venn, Isaac Turner, Iain Mathieson, Natasja de Groot, Ronald Bontrop, and Gil McVean. Strong male bias drives germline mutation in chimpanzees. *Science (New York, NY)*, 344(6189):1272–1275, June 2014.
- [20] Wigard P Kloosterman, Laurent C Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H Moed, Vyacheslav Koval, Ivo Renkens, Markus J van Roosmalen, Pascal Arp, Lennart C Karssen, Bradley P Coe, Robert E Handsaker, Eka D Suchiman, Edwin Cuppen, Djie T Thung, Mitch McVey, Michael C Wendl, Andre Uitterlinden, Cornelia M van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P Eline Slagboom, Dorret I Boomsma, Alexander Schönhuth, Evan E Eichler, Paul I W de Bakker, Kai Ye, and Victor Guryev. Characteristics of de novo structural changes in the human genome. *Genome research*, page gr.185041.114, 2015.
- [21] Laurent C Francioli, Paz P Polak, Amnon Koren, Androniki Menelaou, Sung Chun, Ivo Renkens, Cornelia M van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P Eline Slagboom, Dorret I Boomsma, Kai Ye, Victor Guryev, Peter F Arndt, Wigard P Kloosterman, Paul I W de Bakker, and Shamil R Sunyaev. Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics*, 47(7):822–826, May 2015.
- [22] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernysky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, May 2011.
- [23] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46(8):912–918, July 2014.
- [24] Eric S Lander and Michael S Waterman. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239, April 1988.
- [25] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2):226–232, February 2012.

- [26] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511, October 2002.
- [27] C Aurrecochea, J Brestelli, B P Brunk, J Dommer, S Fischer, B Gajria, X Gao, A Gingle, G Grant, O S Harb, M Heiges, F Innamorato, J Iodice, J C Kissinger, E Kraemer, W Li, J A Miller, V Nayak, C Pennington, D F Pinney, D S Roos, C Ross, C J Stoeckert, C Treatman, and H Wang. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research*, 37(Database):D539–D543, January 2009.
- [28] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.
- [29] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, September 2010.
- [30] Antoine Claessens, William L Hamilton, Mihir Kekre, Thomas D Otto, Adnan Faizullahoy, Julian C Rayner, and Dominic Kwiatkowski. Generation of Antigenic Diversity in *Plasmodium falciparum* by Structured Rearrangement of Var Genes During Mitosis. *PLoS genetics*, 10(12):e1004812, December 2014.
- [31] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert Handsaker, Gerton Lunter, Gabor Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The Variant Call Format and VCFtools. *Bioinformatics (Oxford, England)*, June 2011.
- [32] Alistair Miles, Zamin Iqbal, Paul Vauterin, Richard Pearson, Susana Campino, Michel Theron, Kelda Gould, Daniel Mead, Eleanor Drury, John O’Brien, Valentin Ruano Rubio, Bronwyn MacInnis, Jonathan Mwangi, Upeka Samarakoon, Lisa Ranford-Cartwright, Michael Ferdig, Karen Hayton, Xinzhuang Su, Thomas Wellems, Julian Rayner, Gil McVean, and Dominic

- Kwiatkowski. Genome variation and meiotic recombination in *Plasmodium falciparum*: insights from deep sequencing of genetic crosses. *bioRxiv*, page 024182, August 2015.
- [33] Thomas S Rask, Daniel A Hansen, Thor G Theander, Anders Gorm Pedersen, and Thomas Lavstsen. *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes—divide and conquer. *PLoS computational biology*, 6(9), 2010.
- [34] Susan M Kraemer and Joseph D Smith. A family affair: var genes, PfEMP1 binding, and malaria disease. *Current Opinion in Microbiology*, 9(4):374–380, August 2006.