# 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*

# *List of Figures*

# 1 Motivation

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.[1]

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.[2]

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 respon-

**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 *Aotus nancymaae* | | | | - | + | |

sible. 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 various strains of *P. falciparum* are listed in Table 1.1. Crossing of the pyrimethamine-resistant 3D7 and sensitive HB3 strains[3] revealed a nonsynoymous point mutation in the *dhfr-ts*[1] gene, inhibiting binding of (and thus conferring resistance to) the drug.[4] Analysis of the HB3 x DD2 cross,[5] the latter of which is resistant to chloroquine, localized the determinant to a previously undetected gene on chromosome 7, labelled *pfcrt*[2], 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,[6,7] disruption of which results in marked reduction of mature and functional male gametocytes.

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.[8] These altered forms were likely generated during mitosis by non-allelic homologous recombination[4] (NAHR) of telomeres from two different chromosomes.[9] 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, maladaption). Pathogens are under constant immune and intermittent drug pressure.

[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

NAHR serves to further diversify a pathogen's antigenic repertoire, enabling continued evasion of immunological actors. Random, spontaneous point mutations eventually produce a drug-resistant parasite, capable of surviving the onslaught and reproducing.

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.[10] 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.[?] 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.[?] 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.[11] String matching algorithms that allow mismatches, insertion, and deletions to appear in the alignments allow millions of sequnced 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.[12]

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,[?] the genetic architecture of some common diseases,[13] and mutational rates in humans and chimpanzees with a strong paternal age effect.[?,14–16] Many groups have released sophisticated software packages to facilitate these analyses and detailed instructions on their use.[?,17]

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.

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, DNM could be considered generators of "novel" kmers. 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.

Similarly, if we regard DNM to be generators of novelty, the de novo variant calls we produce with the reference-based protocol should yield an identical set of novel kmers. One must simply take all of the variant calls (inherited and de novo events) obtained for the child and apply this as a delta to the reference genome. At de novo mutations, we extract the local sequence context and divide the data into kmers. If the reference-based analysis and reference-free analysis yield similar number of kmers, then we have found all of the de novo variants.

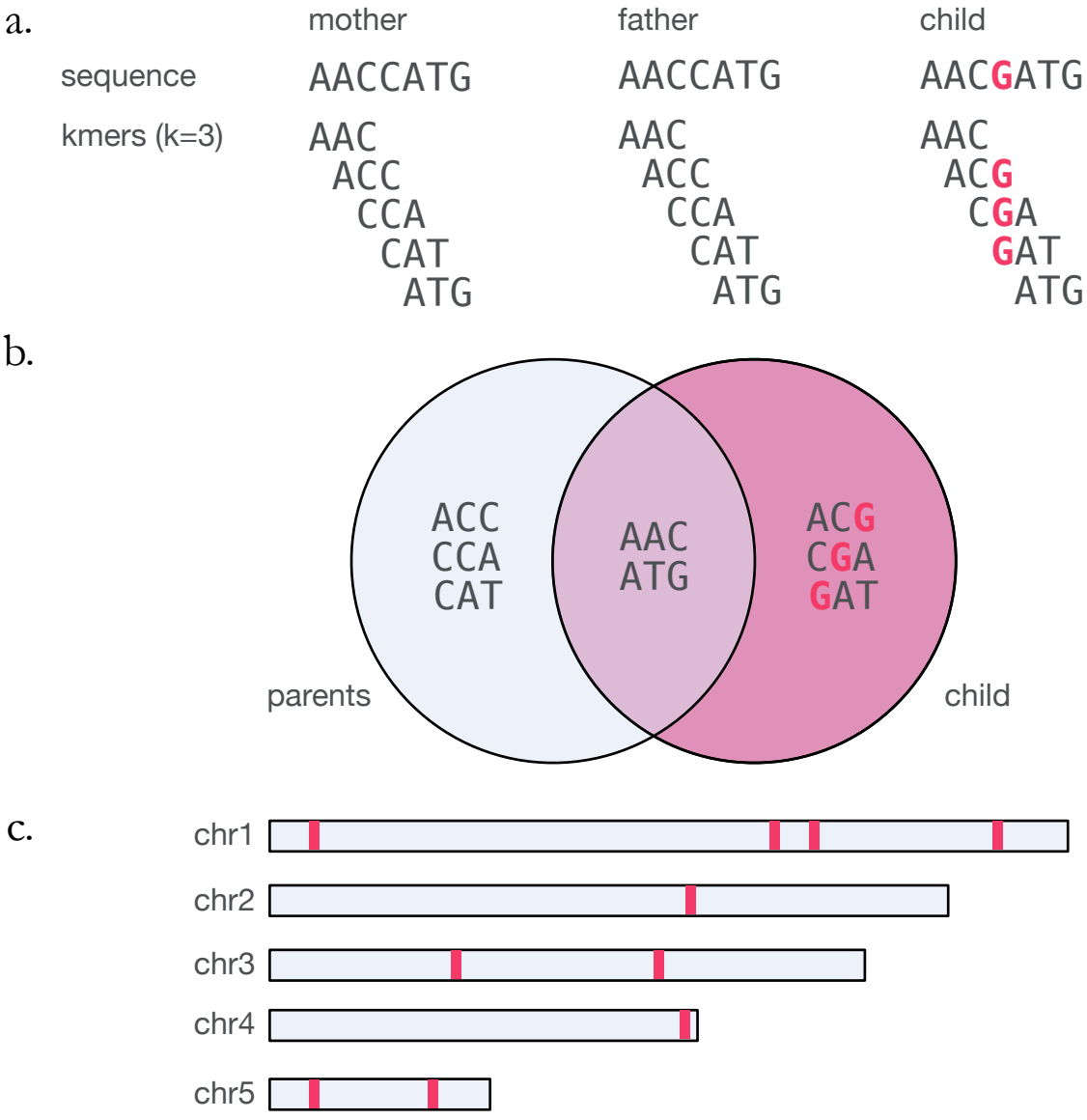


**Figure 1.1:** a. Parental and child sequences at the site of a *de novo* mutation, and the kmers generated at $k = 3$. b. The resulting Venn diagram of kmers found exclusively in the parents, the child, or common to both.

# 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 P.f.

### 2.4.2 Human

### 2.4.3 Chimp

### 2.4.4 Others

# 3 Detection

## 3.1 Basic design

## 3.2 Current state of the art

### 3.2.1 Cross-over

### 3.2.2 Gene conversion

### 3.2.3 Point mutations

### 3.2.4 Structural variation

### 3.2.5 Chromosomal

## 3.3 Potential for de novo assembly to detect events

### 3.3.1 Why assembly should theoretically work

It is conceivable that a mapping-based approach and dozen different variant callers, each purpose-built for dealing with a particular mutational class, could result in a complte callset of variation between samples. However, such a baroque processing pipeline has several drawbacks. In particular, one is beholden to a particular read mapper's determination of a read's unique home in the reference genome. This poses significant problems when reads are of insufficient length to span a repetitive region in the genome, and when the sample being analyzed is sufficiently divergent from the reference such that no plausible home for the read can be found. In these cases, each algorithm is required to go out of its way to adjust read mappings to be more consistent with each other across a locus, or align reads to alternative reference sequences that hopefully contain a closer version of the sequence in question. The result is a hodge-podge of representations of data mapped to a single reference genome in multiple ways, or to multiple reference genomes, callsets with different output formats, etc.

One might consider a different method entirely: *de novo* assembly. In this approach, rather than mapping reads to a reference genome, we dispense with the idea that there should be a single reference genome at all against which variation should be described. Instead, the next-generation sequencing reads are used to produce a "directed acyclic graph" from the data, which is simply a set of vertices (a string representing some amount of genomic sequence) and edges (a set of connections between one vertex and others, representing sequence overlaps). The graph is "directed" in the sense that the connections between vertices are given an orientation that specifies the direction of travel required to obtain linear sequence in the 5′ or 3′ direction. The graph is "acyclic" in the sense that repetitive sequence manifesting as loops will get collapsed into a single loop; the number of times one should traverse the loop in order to recapitulate the input sequence is not stored. Formally, we define this graph as $G = V, E$, a set of vertices and edges between them. The overlaps between reads are detected, and by examining these overlaps, we can walk the graph to detect long contiguous regions of sequence, or "contigs". In theory, with perfect data and long reads that span repetitive regions, the entire linear sequence of a chromosome can be reconstructed in this manner.

Computing the overlaps is a difficult computational challenge to overcome, and there are a variety of methods for doing so. We shall focus on one of the most straightforward methods, effectively amounting to a hashtable approach: construction of a de Bruijn graph. Each read is decomposed into small subwords of a fixed length, $k$, typically referred to as *kmers*. These kmers become the vertices in graph (which, in most implementations, is basically a hashtable). For every kmer in a read, we record the base immediately preceeding and following it. As we process the first (last) kmer in a read, the edge information for the preceeding (following) base is absent. However, if the read overlaps with another one, the overlapping read should contain the same kmer with the missing edge information. We proceed in this manner until all reads are processed and the hashtable is fully populated with every kmer in the dataset and corresponding connection.

For multiple samples, we still construct the graphs separately, but the results can be stored together by introducing the concept of a "color" for each edge. Thus instead of storing a single edge (from a single sample), we store an array of edges, indexed by an integer (or "color") representing each sample. This has the convenient property of allowing us to quickly look up any single kmer in the genomes of the $N$ samples (essentially an $O(1)$ operation) and find edge information for

each sample separately. Variants between the samples will manifest with different edge information for the same kmer, and traversing the graph from where those edges diverge to where they converge again should reproduce the alleles of the variant.

There are tremendous benefits of approaching the problem of finding *de novo* variants this way. First, we are not reliant on the genomes having low divergence from a reference sequence in order to facilitate read mapping. Instead, the reads are effectively aligned to each other, and variants are identified via graph traversals. Thus, in regions of high genomic divergence between samples, we needn't worry that the reads had no place in the reference to map. This gives us a much more complete usage of the data. Second, the graph-based approach is a much cleaner representation of the biology we're attempting to describe. By constructing a "pedigree" graph, that is, a graph representing child, mother, and father, we can quickly identify variants in a child but not in either parent. Furthermore, the graph naturally encodes the parental origin of a *de novo* variant. Should such a variant occur on a haplotype transmitted by the mother, but completely absent in the father, the graph will encode the edge conflict between child and mother and the complete absense of the kmer and edge in the father. This is exceptionally difficult to achieve with a mapping-based approach, even if one tries to map reads to multiple reference genomes and disntangle the results later.

### 3.3.2 *Haploids/perfectly assembled diploids*

Assuming one has perfect data, the assembly of haploid genomes is more-or-less straightforward. The graph is constructed as indicated above. Because the genome is haploid, any kmer with multiple edges in a sample must represent a bit of repeated sequence. Variants will manifest as simple graph bubbles between the samples.

Diploid data is a little trickier, as variants will manifest within a single sample. Should a *de novo* variant occur proximal to a variant between the diploid copies of a chromosome in a sample, it may impair the ...

### 3.3.3 *What do we look for*

...

## 3.4 Limitations of read data

With perfect data, long reads, and uniform coverage, it is theoretically possible to reconstruct entire chromosomes from 5′ to 3′ telomeres. However, data is never perfect, NGS reads tend to be short, and coverage fluctuates.

### 3.4.1 Sequencing errors

Sequencing error can manifest from a variety of sources. Most commonly for Illumina data, error takes the form of single base errors in reads. Rarely, insertion and deletion events may also be present. An understanding of how the sequencing platform works can shed light on these behaviors.

Illumina sequencing is an implementation of so-called "sequencing by synthesis", wherein a small region of the genome is sequenced by providing a single-stranded DNA (ssDNA) template for which the complementary strand must be synthesized. Each nucleotide is labelled with a fluorophore (four different fluorophores for each of the bases, A, C, G, and T) and a "reverseable terminator" which prevents the incorporation of more than one nucleotide per template at a time. The sequencer monitors the synthesis with a camera, taking an exposure each time a nucleotide is incorporated into the fragment. With each cycle of the sequencer, a nucleotide is incorporated, four pictures are taken (each one with a filter that should measure the intensity of flourophores corresponding to a particular nucleotide), and reverseable terminators washed away to prepare for the next cycle. To achieve an adequate signal to noise ratio (SNR), each template will exist in a cluster of thousands of identical templates, and each should incorporate identical nucleotide in identical positions.

Several issues occur in practice. First, nucleotides are labelled with fluorophores that emit light across a range of wavelengths, and the spectra for each overlap slightly. While a nucleotide-specific filter should block out most of the light from the other three, these filters are not perfect, and some light leakage from another set of fluorophores can occur. This "cross-talk" may impair the base-caller's ability to assign the proper label to an event, thus occassionally miscalling nucleotides. Furthermore, the miscalls will be biased towards their cross-talk partners.

Second, with each cycle of the chemistry, a wash step removes the reversable terminator and prepares the flowcell for the next set of labelled nucleotides. However, the wash itself is rather abrasive, and can dislodge templates from the flow-

cell itself. Thus, with each cycle, the clusters of identical templates get smaller and the SNR decays. This can result in diminished ability to call a nucleotide properly.

Third, on occassion, not every template in a cluster will incorporate a nucleotide. More rarely, a template might incorporate two (though the reversable terminator for a nucleotide would have to be absent for this to occur). In either case, subsequent cycles will find some templates out of phase with others in the same cluster. Signal from the cluster is reduced once again, leading to basecalling problems.

Fourth, PCR amplification is used to produce the clusters of identical templates. If an early iteration of the PCR should produce a mismatch, that mismatch will be exponentially propagated to other templates in the cluster, resulting in a consistent error in all of those templates, manifesting as a high-quality single nucleotide mismatch in the read. If PCR amplification is used to amplify the original input DNA as well, this error may be compounded.

Fifth, the wash step that should remove the flourophores from nucleotides in preparation for the next cycle may be inefficient, resulting in a mix of signal from different nucleotides. This effect is clearly observed at the end of homopolymer sequences, where the next nucleotide is often miscalled as whichever base was in the immediately preceeding homopolymer run. Additionally, early versions of the Illumina chemistry (which much of our data will have been sequenced with) suffered from an inefficient wash step for T nucleotide fluorophores, resulting in decreased SNR after Ts are incorporated.

Sixth, inverted repeats can cause single-stranded templates to fold in on themselves, effectively blocking out positions for the labeled nucleotides to anneal. This would force the synthesis to skip ahead several bases, resulting in an apparent deletion in the read.

(TBD: should I talk about error cleaning / correcting in this section?)

### *3.4.2 Read length / repeat length*

Genomes often contain repetitive regions, originating by a variety of mechanisms (tandem duplication, short tandem repeat expansion, etc.). In order to unambiguously reconstruct such regions, read lengths must be long enough to span the repeat and be anchored on either end in unique sequence. In some applications, a paired-end read with sufficiently large insert size might be able to span the repetitive region. Some assemblers (e.g. McCortex) are capable of leveraging such

long-distance information to span repetitive regions and provide information on how many times a graph loop should be traversed. Effectively, this makes the read length one considers when determining the ability to span a repeat to be the mean fragment size of input data, rather than the read length itself.

### 3.4.3 *Coverage fluctuations*

Ideally, a sequencing experiment will yield uniform coverage across the genome. However, in practice this is never the case. Coverage fluctuations occur for a variety of reasons. In particular, high GC content can cause problems during any PCR steps. G and C nucleotides are bound with three hydrogen bonds, rather than the two that form between As and Ts. The ideal temperature to achieve complete denaturation of the strands during amplification may not be precisely known at the time of library preparation, and it may be impractical to measure it for each sample.

Furthermore, if library complexity is low (that is, if the sequencing library is constructed from very little input DNA), an aliquot from this library may suffer from a random sampling bias wherein some sequences are overrepresented (duplicate reads - reads that start that the exact same position in the genome and are effectively multiple copies of the exact same genomic fragment) or underrepresented (little to no coverage over some regions of the genome).

Finally, a genome might not be sequenced to sufficient coverage in order to ensure that all regions of the genome can effectively be sampled. This can occur for certain precious samples where a large amount of DNA cannot be obtained easily and experimenters are forced to proceed with whatever they can get. We can employ Lander-Waterman statistics, which provides estimates for read coverage based on a Poisson model, to compute the amount of genome we might expect to miss when sequencing to a target depth of coverage. Let $a = (N/G) \times L$, where $N$ is the number of reads, $G$ is the genome size (in the case of the *P. falciparum* genome, $\approx$ 23 million bp), and $L$ is the read length (for most of our samples, 76 bp). Then the probability that there are no read starts in some interval $I$ is:

$$p = e^{-a} \tag{3.1}$$

and the probability that there are $\geq 1$ read start in some interval is:

$$q = 1 - e^{-a}. \tag{3.2}$$

```
G = 23e6;
L = 76;

lw = NULL;

for (a in seq(5, 100, by=5)) {
    nReads = a*(G/L);
    nNucleotides = a*G;
    pctGenome = 100 * (1 - exp(-a));

    if (is.null(lw)) {
        lw = c(a, nReads, nNucleotides, pctGenome);
    } else {
        lw = rbind(lw, c(a, nReads, nNucleotides, pctGenome));
    }
}

colnames(lw) = c("coverage", "numReads", "numNucleotides", "pctGenome");

kable(lw, row.names=FALSE);
```

| coverage | numReads | numNucleotides | pctGenome |
|---|---|---|---|
| 5 | 1513158 | 1.150e+08 | 99.32621 |
| 10 | 3026316 | 2.300e+08 | 99.99546 |
| 15 | 4539474 | 3.450e+08 | 99.99997 |
| 20 | 6052632 | 4.600e+08 | 100.00000 |
| 25 | 7565789 | 5.750e+08 | 100.00000 |
| 30 | 9078947 | 6.900e+08 | 100.00000 |
| 35 | 10592105 | 8.050e+08 | 100.00000 |
| 40 | 12105263 | 9.200e+08 | 100.00000 |
| 45 | 13618421 | 1.035e+09 | 100.00000 |
| 50 | 15131579 | 1.150e+09 | 100.00000 |
| 55 | 16644737 | 1.265e+09 | 100.00000 |
| 60 | 18157895 | 1.380e+09 | 100.00000 |
| 65 | 19671053 | 1.495e+09 | 100.00000 |
| 70 | 21184211 | 1.610e+09 | 100.00000 |
| 75 | 22697368 | 1.725e+09 | 100.00000 |
| 80 | 24210526 | 1.840e+09 | 100.00000 |
| 85 | 25723684 | 1.955e+09 | 100.00000 |
| 90 | 27236842 | 2.070e+09 | 100.00000 |
| 95 | 28750000 | 2.185e+09 | 100.00000 |
| 100 | 30263158 | 2.300e+09 | 100.00000 |

(Note: Surprising - I would have expected the need for a higher coverage to capture the genome fully... double-check these results)

### 3.4.4 *Rare vs error*

...

### 3.4.5 *Algorithm error can mimic real biology*

Finally, the assembly algorithm employed might make errors that mimic real biology, confounding the results. For example, an overzealous error-correction algorithm may attempt to correct a low-frequency nuclotide in the read data under the assumption that its sparsity reflects an error process, when it fact it is derived from a repetitive sequence that has slightly deviated from the original ancestral sequence, and simultaneously has poor coverage that masks the problem. This would hide the divergence of one of the repeat's copies.

Additionally, in an error-cleaning step that should look for short branches in the graph that are likely derived from sequencing error and discard them, the algorithm may "over-clean" (too many events are thrown away) or "under-clean" (too many events are retained). These may conspire to artificially reduce

effective contig length by either removing critical kmers or presenting too many ambiguities in the graph to effectively complete the traversal.

Finally, some assembly algorithms will attempt to leverage long-range information to scaffold short contigs into longer "supercontigs". This relies on the ability to map the original paired-end reads back to the contigs and detect read pairs anchored in two contigs. Should the alignment be in error (due to aforementioned sequencing errors, coverage fluctuations, etc.), the scaffolder may join two contigs that do not represent a real sequence in the genome. This may manifest as a structural variant or recombination event where none has occured.

## *3.5 Outline of the work*

# 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 interchromosomal 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, $k = 47$).

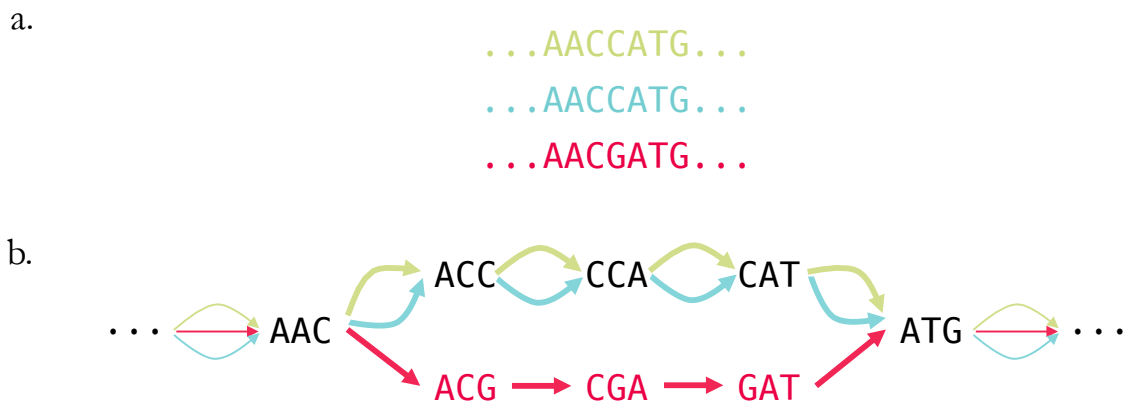


**Figure 4.1:** a. Haploid sequences from a mother (green), father (blue), and child (red), the last differing from the first two by a single SNP. b. The resulting multi-color de Bruijn graph for $k = 3$. Red vertices denote kmers that are deemed "novel", i.e. present in the child and absent in the parents. 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.

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).

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

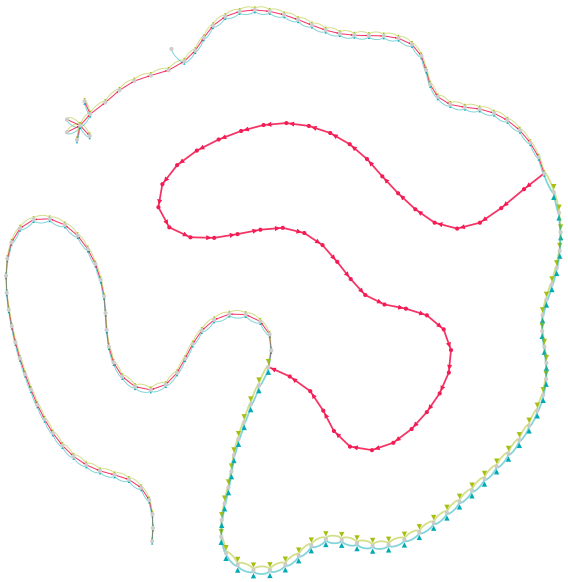

**Figure 4.2:** A multi-color de Bruijn graph at $k = 47$ for a haploid pedigree spanning a simulated *de novo* SNP. Vertex labels have been supressed for clarity. Spatial layout is arbitrary and for display purposes only.

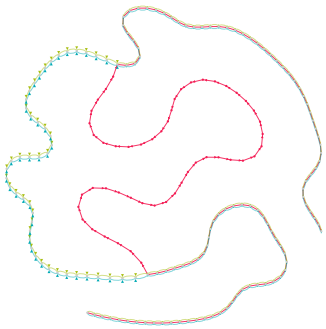

**Figure 4.3:** A 5 bp insertion in the child

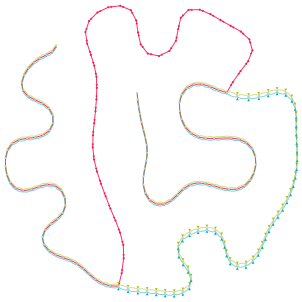

**Figure 4.4:** A 5 bp deletion in the child

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.

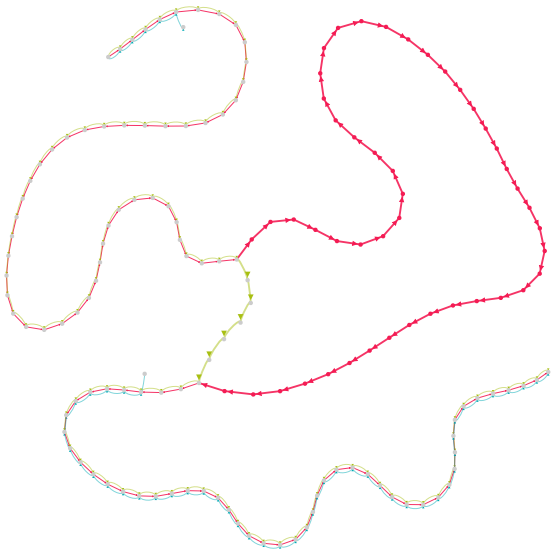

**Figure 4.5:** A tandem duplication on the haplotypic background of the mother.

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*

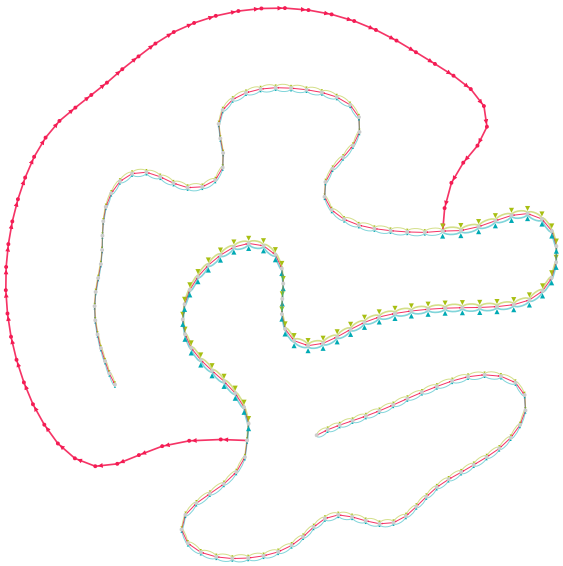

**Figure 4.6:** A variant wherein the child's path does not simply diverge from that of the parents, but rather navigates both.

# 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] William Bateson and Miss E R Saunders. Experimental Studies in the Physiology of Heredity, 1908.

[2] T H Morgan. Sex Limited Inheritance in Drosophila. *Science (New York, NY)*, 32(812):120–122, July 1910.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(11s):S6–S12, November 2009.

[12] 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.

[13] 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, Seungtai 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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 Kernytsky, 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.