Methods for Quantifying Systematic Read Mapping Errors in DNA Resequencing Techniques

by
Peter Georgeson

Supervisors:
Professor Justin Zobel
Dr Jan Schröder

Submitted in partial fulfilment of the requirements for the degree of
Master of Science (Computer Science) at the University of Melbourne, Australia.

November 2, 2015

Student Number: 659065
Project Type: Research Project
Credit Points: 75cp
Subject Code: COMP60003

Declaration

I certify that

- this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text.
- where necessary I have received clearance for this research from the University's Ethics Committee and have submitted all required data to the Department.
- the thesis is 20,258 words in length (excluding text in images, tables, bibliographies, and appendices).

Peter Georgeson
Computing and Information Systems
University of Melbourne, Australia
November 2, 2015

Acknowledgements

I am grateful to my supervisors, Professor Justin Zobel and Dr Jan Schröder for their dedicated support and encouragement throughout this project. Justin and Jan's insights and guidance, along with their patience and commitment, have been instrumental in the completion of this thesis. I am also grateful for the advice provided by several colleagues. In particular, I thank Dr Bernie Pope for his feedback and suggestions. Finally, my family, and in particular, my wife, have always been supportive and understanding. They make this journey worthwhile.

Contents

List of Figures

List of Tables

Abstract

Recent advances in genome sequencing have transformed biomedical research by enabling the generation of billions of short DNA fragments covering an individual's genome. A critical challenge is to use these fragments to reconstruct the individual's genome. A common approach to this reconstruction is to align sequence fragments, called reads, to a representative reference genome, with the goal of identifying differences between the individual and the reference. From these differences the genome of the sequenced individual can be determined.

Reference bias arises when the choice of alignment target unduly influences the inferred genome of the individual. For example, regions of similarity between the individual and the reference genome tend to map successfully, while regions of significant difference may be lost. Consequently, this generates systematically biased results, a problem that increases as differences between the individual and the reference accumulate. Despite this, the causes, prevalence, and effect of reference bias are not well characterised.

Using both simulated data and real data, we identify the factors leading to reference bias, and measure its impact on results. We find that even if an individual's genome exactly matches the reference genome, some regions cannot be aligned confidently. The prevalence of repeated regions varies significantly across genomes, with variability observed even within the same species.

We demonstrate that current methods can robustly detect small variations, provided that mutation rates remain below 4%. However, at one location in 10,000, small variations cannot be detected, while at one location in 1000, the ability to detect a variation is significantly impaired. We find two primary causes of reference bias when predicting short variations: short tandem repeats and similar sequence. Read length determines the length of variation that can be directly detected, while sequencing coverage enables more variations to be confidently predicted.

By utilising whole-genome alignment, we measure reference bias across the *E. coli* species, finding that between 0.2% and 29.9% of an individual's genome is unrecoverable, depending on the choice of reference. Sequence from the individual's genome that is not present in the reference genome is a major factor that limits accurate reconstruction of the individual's genome, and is a significant obstacle to accurately recovering an individual's DNA sequence using alignment alone.

Chapter 1

Introduction

With the massive growth in volume and availability of genomic data, it is becoming feasible to investigate questions such as the impact of differences between genomes. Genetic variants have been associated with many human disorders, including cancer, diabetes, and cardiovascular disease, as well as many other traits. Understanding the genetic basis of disease can improve treatment and prevention.

Sequencing data can help to provide this understanding. Our capacity to sequence the DNA of organisms has continued to grow, due in part to the adoption of high-throughput sequencing (HTS), also known as next-generation sequencing (Schuster, 2008). HTS has increased throughput and reduced the cost of sequencing by several orders of magnitude in less than ten years (Isakov & Shomron, 2011). Between 2001 and 2015, the average cost per megabase of DNA sequence dropped from approximately $US5,000 to $US0.05 (Wetterstrand, 2015). The result is a revolution in the field of genomic analysis.

When a representative reference genome exists for an organism, resequencing can be employed. Short reads are aligned to the reference genome by searching for segments in the reference that match each read. The aligned reads enable differences to the reference to be inferred, and from this the complete genome can be reconstructed. Alignment is significantly less demanding than assembly: resequencing requires far less time, data, and computing resources. Resequencing has become increasingly common as more reference genomes have become available.

A potential disadvantage of resequencing is that the reconstructed genome can be unduly influenced by the reference. In particular, reads that do not closely match any region on the reference may be discarded, resulting in a systematic bias favouring reads that closely match sequence on the reference. The process of generating reads destroys information linking the read to its original location. Hence, any similar region across the entire reference is a potential location for a read. This ambiguity exacerbates the problem.

As a result, reference bias impacts on subsequent analysis by systematically preferring specific alignments. Evidence for particular types of variation may be greatly reduced or completely eliminated. Worryingly, alignment results can appear sensible even when significant variations have been lost during the alignment process.

Although reference bias is known to be a factor that affects the accurate determination of variations based on alignment, there is limited understanding of its causes and effects, or of the specific circumstances where it requires consideration. In short, the severity, extent, and impact of reference bias is largely unknown.

This is independent of the fact that sequencing is subject to many biases prior to alignment, such as GC-content (Benjamini & Speed, 2012), read position (Schwartz et al., 2011), and sequence motifs (Allhoff et al., 2013). In this study, we only consider the effect of *reference bias*. Specifically, reference bias refers to systematic bias or errors arising from the structure and content of the reference genome.

Given that reference bias is a poorly understood confounder affecting alignment-based genomic analysis, we aim to quantify the effects of reference bias by identifying the circumstances in which it is a significant factor. This aim gives rise to the following goals of this research:

- *Identify scenarios where reference bias is an issue:* by quantifying reference bias, we demonstrate circumstances in which reference bias arises. We provide methods of assessing the likely impact of reference bias on a specific experiment.
- *Recommend approaches to reducing reference bias:* we outline and assess approaches to limiting reference bias.
- *Limits to alignment:* we demonstrate the limits to variation discovery using alignment.

We take a systematic approach to the identification of reference bias, with the initial analysis focused on a simulation framework. With complete control over the reference, the included variations, and all other parameters of the experiment, we generate instances illustrating reference bias. In this simulation environment, we measure the impact of reference bias on downstream analyses, and identify the precise causes of reference bias.

Multiple reference genomes are now available for numerous organisms. We analyse and measure the effect of the choice of reference using existing datasets. This enables inferences to be made about the likely impact of the choice of reference, including circumstances where only one candidate reference is available for a given genome.

An important factor influencing the impact of bias is the purpose of the analysis, or what biological knowledge is being sought. Initial experiments focus on single nucleotide variations as the simplest form of variation. This type of variation is also the simplest to detect and the most robust against the effects of reference bias. The framework is expanded to measure the impact on other types of variation, including short insertions and deletions, and, finally, larger structural variations. This enables us to measure the impact of bias on different types of variation and, critically, we illustrate the significant variability of the effect of reference bias, depending on the parameters of the experiment. We show that, in many circumstances, reference bias has negligible impact, while in others reference bias severely limits the ability to obtain any sensible results.

Chapter 2

Background

In this chapter we provide a brief introduction to molecular biology and describe current methods of determining an organism's genome. We identify existing issues with these approaches with the aim of providing a context for the issue of reference bias.

2.1 Biology

Life consists of cells. Cells have the ability to regulate their own behaviour, using a set of instructions that, among other things, specify the assembly of proteins. These instructions are generated from long, sequential chains of molecules called *deoxyribonucleic acid* (DNA).

DNA consists of a linear sequence of nucleotides. There are four types of nucleotide: adenine, cytosine, guanine, and thymine, denoted as A, C, G, and T respectively. Hence DNA can be represented as a string consisting of the alphabet A, C, G, and T. Each distinct sequence of DNA is a *chromosome*, and the collection of all DNA sequences in a cell is its *genome*.

DNA molecules are usually double-stranded, consisting of two bonded strands that form a double helix. The sequence of one strand can be derived from the other because of the complementary nature of DNA: adenine is paired with thymine and cytosine is paired with guanine. Consequently, each element in a DNA sequence is often referred to as a *base pair* (bp). Strands of DNA are read in opposite directions: each strand is the *reverse complement* of the other. This redundancy of information provides the mechanism for DNA replication. In addition to being double stranded, some organisms have multiple copies of each chromosome. Organisms with one, two, or more than two copies of each chromosome are referred to as *haploid*, *diploid*, or *polyploid* respectively. Each copy is referred to as a *haplotype*. If a variant occurs in both haplotypes, it is *homozygous*; if it occurs in just one haplotype, it is *heterozygous*.

To generate proteins, subsequences of the DNA are copied into RNA, a step referred to as *transcription*. Parts of this transcript can be spliced out, before they are *translated* into a protein. The proteins expressed in a cell are responsible for nearly every task that a cell performs.

When genetic material is transferred from one generation to the next, differences accumulate.

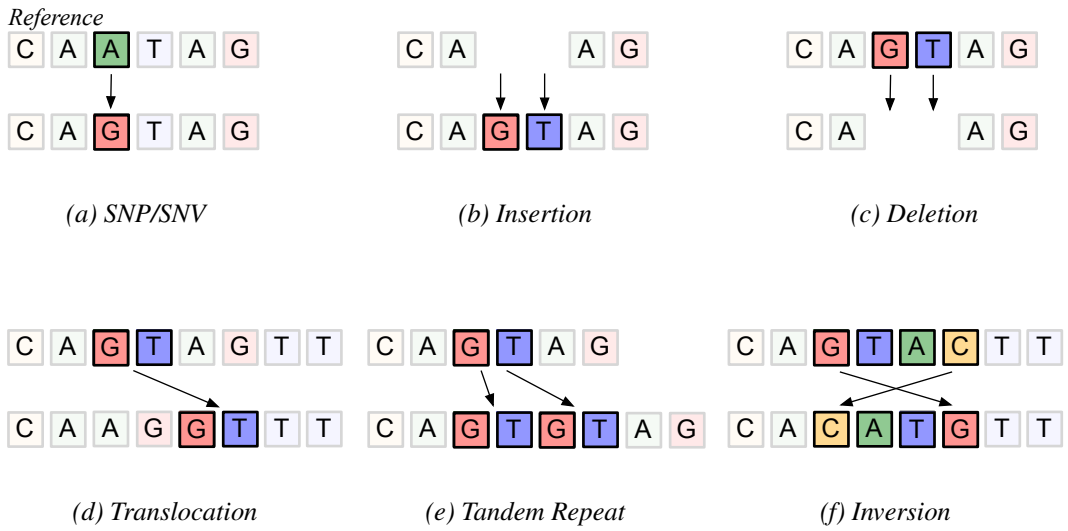

Figure 2.1: Variation types. Variations are described by the operation that is performed on the reference to produce the individual's DNA sequence.

These differences define the unique characteristics of each individual: a variation in a region that encodes a protein can directly affect its expression, while changes in non-coding regions can alter the level of expression. Other changes may carry no discernible effect. The observable characteristics of an organism are collectively referred to as the organism's *phenotype*.

Variations take several specific forms and are illustrated in Figure 2.1. Most variations are *single-nucleotide variations* (SNVs): the mutation of a single base. SNVs that are considered common to a population are referred to as *single-nucleotide polymorphisms* (SNPs). The 1000 Genomes Project Consortium et al. (2010) found 15 million SNPs in a population of 179 individuals, indicating that SNPs cover approximately 0.5% of the human genome.

Short insertions and deletions (*indels*) are also common. In this instance genetic material is inserted into or removed from the sequence. Indels are functionally important, and have been implicated in numerous diseases (Miki et al., 1994; Ball et al., 2005). Proteins are encoded by reading triplets of nucleotides; unless the indel size is a multiple of three, the subsequent frameshift can significantly alter protein expression by affecting all downstream transcription (Gonzalez et al., 2007). As a result, indels are an important component of the study of genetic causes of disease.

Although estimates of indel frequency vary considerably, indels are clearly a common variation, and are the most common after SNVs. Cartwright (2009) estimates an indel frequency of approximately 14% relative to SNVs, while Mills et al. (2006) found an average density of one indel every 7.2 kbp, estimating that indels are responsible for 16%–25% of all genomic variation in humans.

Short tandem repeats (STRs) are a specific type of indel. STRs are directly adjacent repeating sequences of identical content, usually with a period of less than 10 bp. STR variations have numerous genetic applications, including DNA fingerprinting (Kayser & de Knijff, 2011), and have been implicated in several genetic disorders, such as Huntington disease (Pearson et al., 2005).

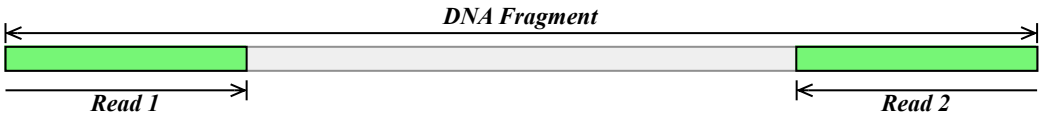

Figure 2.2: Paired-end reads. Reads are generated from each end of a DNA fragment, providing additional information about the distance between the two reads.

Large-scale variations that are not localised to a small region of the genome are referred to as *structural variations* (SVs). An SV is a genomic rearrangement affecting more than 50 bp of sequence (Alkan et al., 2011); this arbitrary cutoff is sometimes defined as 1000 bp (Medvedev et al., 2009). SVs include long insertions, long deletions, inversions, duplications, and translocations (see Figure 2.1). These variations can also be combined to form complex SVs.

SVs are increasingly recognised as a significant component of genetic diversity. SVs have been implicated in genetic disorders and are a hallmark of cancer genomes (Bashir et al., 2008; Lee et al., 2010; Hillmer et al., 2011). Recent studies suggest that SVs are frequent in the human genome. Zhang et al. (2009) found that SVs cover more than 30% of the human genome, demonstrating the tremendous variation they generate.

DNA contains many artifacts as a result of evolution. A significant proportion of repetitive content in DNA derives from *transposable elements* (TEs). TEs are mobile DNA sequences that migrate to different regions of the genomes, often leaving copies of themselves behind. Similarly, *pseudogenes* are copies of existing genes that have become non-functional, while *paralogs* are related genes that have diverged. Importantly, these evolutionary mechanisms result in regions of similar sequence across a genome.

Cellular life can be divided into two domains: *prokaryotes* and *eukaryotes*. Prokaryotes include bacteria and are always unicellular. Eukaryotes include the animal kingdom, and have a more complicated cell structure. Fundamental differences in the cell structure of these domains has implications for the DNA content that is found in each (Vellai & Vida, 1999).

2.2 High-Throughput Sequencing

Sequencing is the process of determining the linear order of nucleotides in a DNA fragment. This is primarily achieved using the process of *shotgun sequencing*. DNA molecules are cut into fragments, then these fragments are copied millions of times in a process called *amplification*. Fragments are then *sequenced*: bases are read from the end of each fragment to generate a *short read*, with read lengths typically in the range of 100 bp to 1000 bp. *Single-ended reads* are taken from one end of each fragment; *paired-end* and *mate-pair* techniques provide pairs of reads that have been sequenced from each end of a fragment (see Figure 2.2).

Demand for low-cost sequencing has driven the development of HTS techniques. These tech-

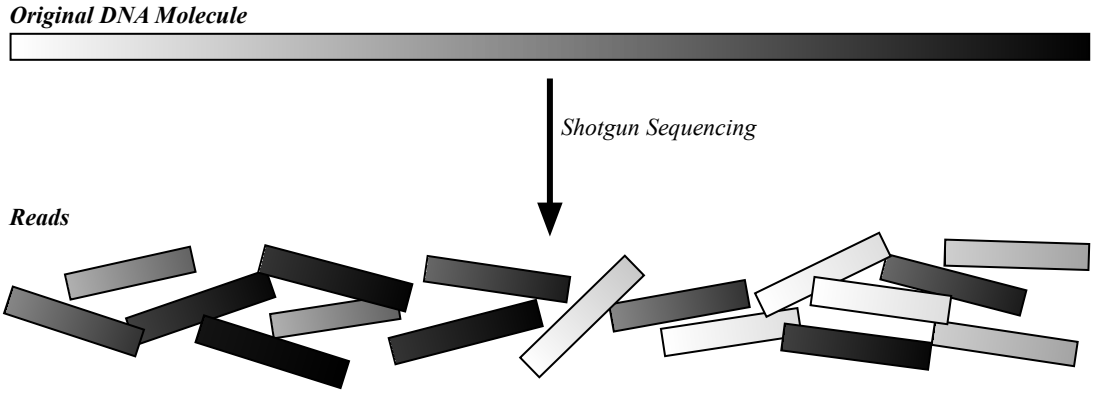

Figure 2.3: Shotgun sequencing. The DNA is broken up into millions of small fragments before it can be sequenced.

niques parallelise the sequencing process to dramatically reduce cost while greatly increasing the number of reads generated (Shendure & Ji, 2008). However, read lengths are typically just a few hundred bases, and error rates are relatively high, of the order of 1%.

Reads can originate from either strand of DNA, and, in diploid organisms such as humans, from either chromosome. An implication of this process is that all information about the location of each fragment in the original piece of DNA is lost (see Figure 2.3). Inference of the original DNA sequence from the set of generated reads is a computationally difficult task.

In this study, we primarily consider single-ended (unpaired) reads originating from haploid organisms. Paired reads are typically 300 bp to 500 bp apart, though the exact width separating each pair is not known. Subsequently, paired reads can be considered to be a long single-ended read with a length spanning the fragment, but containing a region of unknown content and uncertain width.

2.3 Resequencing

When a genome has no existing reference, or is expected to have significant structural variation, *de novo assembly* is employed. Assembly proceeds by searching for overlaps in the generated reads. Overlapping reads are combined to form a graph, with each path through the graph representing a plausible subsequence in the original sequence (Compeau et al., 2011). Reconstructing a complete genome using de novo assembly remains a challenging problem for larger genomes. The output of an assembler is typically a large number of contiguous sequences rather than a complete genome.

If a genome that is representative of a species is available, an alternative to assembly is *resequencing*. Reads from an individual, the *donor*, are aligned to a consensus genome: the *reference genome*. Given a reference genome and a set of reads from a donor genome, each read is independently aligned to the region that is most similar to that read (see Figure 2.4). A *pileup* of reads at each genome position can then be analysed to determine if a difference between the donor and the reference exists

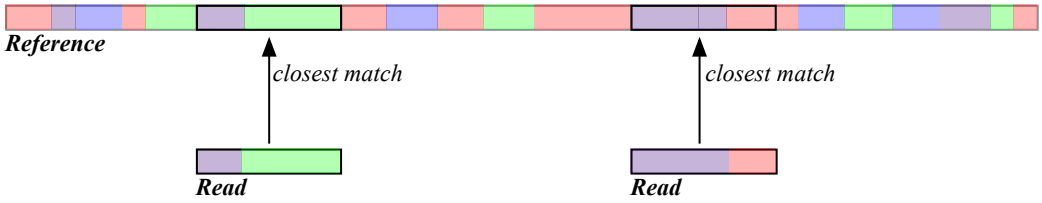

Figure 2.4: Aligning reads to a reference genome. Alignment software considers each read independently and finds the closest matching location on the reference genome.

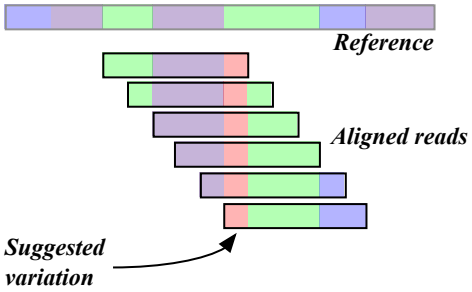

Figure 2.5: Inferring variations using alignment: when a sufficient number of aligned reads report the same difference at a specific location, a variation between the donor genome and the reference is suggested.

(see Figure 2.5).

If the read contains errors, the donor differs from the reference, or the read falls in a repeated region, then the read will not uniquely and exactly match the correct location on the reference. Finding the correct location for a read thus becomes an *approximate string matching problem*.

The Smith-Waterman algorithm (Smith & Waterman, 1981) enables similar regions to be defined in terms of *edit distance*. The edit distance between a read and a potential location is defined by the set of edits – substitutions, insertions, or deletions – that are required to transform the read into the proposed location on the reference.

Each type of edit is assigned a cost, and the algorithm finds one or more locations on the genome with the minimum cost. The cost of each type of edit can be modified to more accurately model the expected types of read errors, and the biological likelihood of each type of change. Insertion and deletion costs are usually generalised to gaps. Biologically, one large insertion is more probable than numerous smaller insertions, so an affine gap penalty is also typically introduced in the form of "gap open" and "gap extension" penalties. Although this is a more realistic model, long gaps actually occur *more* frequently than short gaps. Using the edit distance paradigm to model sequences has limitations.

Although the Smith-Waterman algorithm finds locations with the lowest edit distance, its time complexity for each read is $O(mn)$, m being the length of the read and n being the length of the genome. This is infeasible for aligning billions of reads to genomes of billions of bases. To achieve

	Match	Mismatch	Gap Open	Gap Extend
BWA SW	+2	-5	-2	-1
BWA MEM	+1	-4	-6	-1
Bowtie 2	+2	-2 to -6	-5	-3
LAST (long reads)	+1	-1	-7	-1
LAST (short reads)	+6	-18	-21	-9
Subread	+2	0	-2	0

Table 2.1: Default scoring systems of popular aligners

acceptable time complexity, aligners typically pursue one of three strategies: hash tables, FM-indexing (Ferragina et al., 2009), or merge sorting.

To deal with approximate matches, hash-based aligners employ a seed-and-extend strategy. This is achieved by considering all distinct substrings of fixed length k (referred to as *k-mers*). By computing hashes of all k-mers in the reference, and all k-mers in the read, potential matches can efficiently be found. The matching seed is extended using a dynamic programming algorithm to evaluate its suitability as a matching location. Similarly, FM-index based schemes find the longest exact match covering each position in the read, then find and evaluate approximate matches by extending these seeds. In general, hash-based aligners are more sensitive, but tend to be slower and require more memory, relative to FM-index based aligners (Lee et al., 2014).

The implementation of these algorithmic shortcuts and heuristics has resulted in the development of several efficient alignment programs (Li & Homer, 2010), including *BWA* (Li & Durbin, 2010), *Bowtie 2* (Langmead & Salzberg, 2012), *LAST* (Frith et al., 2010), and *Subread* (Liao et al., 2013). Billions of reads can be aligned to large genomes such as the human genome in a few hours, on conventional hardware. However, there has been little systematic analysis of the relationship between the accuracy of the alignment, the effect of edit distance parameters, and the various alignment heuristics. Table 2.1 demonstrates the lack of consensus with respect to edit distance penalty weights for a selection of popular alignment tools. Optimal edit distance parameters are dependent on the nature of the problem, and preferred parameters are often selected through trial and error (Quick et al., 2014).

Adding to the difficulty of assessing different aligners and selecting the best configuration is the challenge of obtaining data sets with known correct alignments. One solution is to generate simulated data. The generation of a synthetic donor genome enables known variants to be added, while synthetically generated reads can have known errors added. This is a common approach: several surveys evaluating alignment accuracy and variant prediction have been conducted using simulation-based approaches (Schbath et al., 2012; Li et al., 2008), and simulation tools such as *wgSim* (Li, 2013) and *Seal* (Ruffalo et al., 2011). A disadvantage of simulated data is that it measures which aligner most closely matches the parameters that generated the data, which does not necessarily correspond to real data. It is difficult to accurately simulate artifacts such as the non-random distribution of variants and the dependent nature of errors (Li, 2014).

Evaluating real data has its own set of challenges. One approach is to compare the results of competing toolsets to one another, but this biases results towards the consensus result. In particular, this method does not discover systematic errors. Another approach is to use results collected from a different technology. Similarly, this approach does not discover systematic biases. This method also assumes that the alternative result set is at least as accurate as the technology being assessed. As a result, there is little consensus on variation calling error rates or on true variation rates. For example, estimates of indel frequency vary considerably. Although several tools exist for predicting indels (Li et al., 2009a; Albers et al., 2011; Krawitz et al., 2010), recent analyses demonstrate significant differences in results (Pabinger et al., 2014; Neuman et al., 2013).

The strategy of aligning reads to a reference to infer variations and ultimately the donor sequence has clear advantages, primarily that this solution is computationally tractable for large genomes with substantial read coverage. However, the underlying assumption of this strategy is that most reads align correctly. Systematic reference bias arises when this assumption fails to hold as a result of the reference sequence. Intuitively, the likelihood that a read will align correctly is proportional to how closely it matches its correct location, and is inversely proportional to how closely it matches other incorrect locations. Trivially, a read that exactly and uniquely matches its correct location will always align correctly. In contrast, a read that contains errors, varies significantly from the reference, or aligns equally well to multiple locations will tend not to be aligned, will be aligned with low confidence, or will be aligned incorrectly. As a result, areas in a reference genome containing repeated sequence are less likely to be aligned to and variations in these regions are less likely to be discovered.

We outline the causes, impact, and existing solutions to the problem of reference bias in the next chapter.

Chapter 3

Reference Bias

We have outlined the alignment process, describing how reads from a donor genome that are aligned to a reference genome can be analysed to find variations, ultimately enabling the inference of a consensus sequence that represents the donor genome.

The assumption of this process is that most reads align correctly, but there are various circumstances where this assumption breaks down. For instance, areas where the donor genome and reference genome diverge are less likely to align correctly, if at all.

In this chapter we describe specific examples of the impact of reference bias, before outlining current solutions to this issue. We then identify the causes of reference bias and suggest methods for measuring reference bias.

3.1 Consequences of Reference Bias

Reference bias arises in a range of circumstances:

- In any population of organisms, there is genetic diversity across the population. If there is a single reference genome, naturally there will be some individuals who are genetically relatively similar to this reference, and other individuals who are relatively distant from this particular sample. For example, the donors used to create the reference human genome primarily lived in Buffalo, New York and were likely to be white Caucasians (Osoegawa et al., 2001). Consequently, resequencing a non-Caucasian individual introduces reference bias as they vary more from the reference genome. This effect is present in any population that has significant genetic diversity.

- Cancer cells typically exhibit far higher levels of mutation than non-cancer cells. Cancer is caused by genetic instability and is characterised by rapid rates of mutation. Coupled with the difficulty of obtaining cancer cells containing the same set of mutations, resequencing cancer cells remains a challenging task, due partly to the problem of reference bias.

- The structure of the genome has a significant impact on the accuracy of resequencing. Repeated sequences are a major impediment to accurate alignment (Trapnell & Salzberg, 2009). At least 50% of the human genome consists of repeated sequences (Lander et al., 2001), while 84% of the maize genome is covered by transposable elements (Schnable et al., 2009). Reference bias is more of a problem if the genome is already difficult to align to due to artifacts such as long repeated sequences.

Some applications of aligned data are particularly sensitive to reference bias. For example, experiments that infer results based on coverage or counts of aligned reads assume that reads are uniformly sampled and uniformly aligned to the reference genome. This assumption is incorrect unless reference bias is taken into account. Specific examples illustrating the impact of reference bias in this context are described below.

- *Allele-specific expression (ASE):* ASE is a mechanism that results in the two alleles in a diploid individual being expressed at different rates. ASE is a source of biologically interesting phenomena such as genomic imprinting (Babak et al., 2008). Alleles associated with this mechanism can be discovered using resequencing. Expressed transcripts are sequenced and aligned to a reference genome. The areas of interest are restricted to sites containing heterozygous point mutations, then the relative counts of aligned reads containing the reference allele and aligned reads containing the alternative allele are compared. If the counts of each allele differ significantly, then the presence of ASE can be inferred.

  Reads containing the reference allele are guaranteed to match the reference genome at the position of the allele, whereas reads containing an alternative allele are guaranteed of a mismatch at the position of the allele. This additional mismatch increases the likelihood that the read will fail to align to the reference, resulting in a bias against reads containing non-reference alleles. Degner et al. (2009) observed that a significant proportion of alleles appearing to demonstrate strong levels of ASE were actually false positives that were caused by reference bias. In general, reference bias results in an underrepresentation of reads matching the non-reference allele.

- *Phylogenetics:* A common application of HTS is the reconstruction of evolutionary trees. Reads from each individual are aligned to a single reference genome. Differences in SNP alignments are used to infer evolutionary distances, and ultimately a phylogenetic tree for the population. Bertels et al. (2014) illustrate that the choice of reference genome can significantly affect the final phylogenetic tree, particularly within diverse populations.

- *Copy number variations (CNVs):* Large-scale duplication of DNA sequence, whether applied to a specific gene or an entire chromosome, is predominantly detrimental. CNVs are associated with numerous diseases, including Down syndrome, psychiatric disorders, kidney defects, and heart disease (Tang & Amon, 2013).

  A common method of detecting CNVs is with depth of coverage data. Once reads are aligned to the reference genome, coverage depth is determined by counting the number of reads covering

each position. Duplicated sequence results in increased coverage in the region that has been duplicated, which is then used to infer copy number counts.

Methods that rely directly on read depth coverage are particularly prone to the effects of bias. Since read depth is directly affected by GC-content and repeated sequence, many tools correct for this bias (Lai & Ha, 2014). In repeated regions, the number of reads that align to a region is no longer strongly correlated with the number of reads sequenced from that region. As a result, sensitivity in repeated regions is typically drastically reduced (Teo et al., 2012). The alignment strategy influences whether deletions or insertions tend to be misreported (see Section 3.3.2), while simply excluding repeated regions substantially increases false negatives, particularly as many SVs lie in repeated regions (Medvedev et al., 2009).

- *Metagenomics:* Metagenomic sequencing reads are derived directly from an environmental sample, such as soil, seawater, or the gut. A typical goal is to determine which organisms are present in the sample, and their abundance. This is often achieved by aligning reads to a collection of reference genomes, with abundance determined by analysing the number of reads aligning to each genome.

  Issues of bias are magnified when applied to metagenomics experiments, with results being extremely sensitive to the selection of reference genomes chosen to align against. A metagenomics study of bacteria present on the New York subway indicated the presence of anthrax and bubonic plague (Afshinnekoo et al., 2015), but spurious alignments caused by contamination are common (Lusk, 2014). Strains of bacterial species tend to have very similar sequences. Consequently, any minor variations in the generated reads, alignment process, or reference sequence can cause major differences in the final results (Reinert et al., 2015).

- *Transcriptome alignment:* Failure to carefully consider the level of repetition and near-repetition in the genome can lead to spurious results. For example, Li et al. (2011) found unexpectedly high levels of variation in RNA transcripts when aligned and compared to the reference genome, generating speculation that a new biological mechanism for RNA editing had been discovered. More detailed analysis by Schrider et al. (2011) revealed that reference bias had caused many false positives, primarily due to paralogous genes resulting in repeated sequences, and aligner limitations at regions of high variability.

3.2 Current Solutions

Existing methods of reducing the effect of the reference fall into the following categories:

- *Statistical methods*: by applying improved statistical models and incorporating all available data, more robust inferences can be made.

- *Filtering*: the genome is analysed for areas that are difficult to align or are prone to bias. These areas are excluded from the analysis.
- *Alignment strategies*: alignment software is configured specifically to address bias.
- *Reference augmentation*: existing catalogs of known variations are incorporated into the alignment process to improve their discovery.
- *Multiple genomes*: rather than align to a single reference genome, a family of genomes is used as an alignment target.

Specific examples of these approaches are outlined below.

3.2.1 Statistical Methods

Early methods of predicting variants used simple thresholds to select the type of variation (Harismendy et al., 2009; Nielsen et al., 2011), but this method is only effective at high levels of coverage. Consequently, recent strategies include probabilistic methods that assign prior probabilities to events. Probabilistic frameworks enable errors and biases to be incorporated into the detection model, which enables statistically robust predictions to be made.

Probabilistic techniques enable a confidence level for each genotype to be provided (Li et al., 2009b). Tools such as *Crossbow* (Langmead et al., 2009) incorporate databases of known variations into the probability model to improve likelihood estimations. Similarly, Skelly et al. (2011) introduce Bayesian priors into the ASE detection process. This enables technical variation to be included in the model and can significantly reduce the incidence of false positives due to reference bias.

These techniques maximise the use of available information, which provides more accurate predictions and statistically interpretable confidence levels, but if information has already been lost during read alignment it cannot be recovered. For example, if no reads map to a highly divergent region that contains a SNV, downstream tools cannot predict the variant, since the alignment contains no information about that region.

3.2.2 Filtering

An approach to improving specificity is to exclude regions of the genome that are unlikely to produce correct alignments. The concept of *mappability* as a function of location has been explored in detail (Derrien et al., 2012; Lee & Schatz, 2012). Mappability measures the uniqueness of a sequence within a genome, and is a strong predictor of the depth of aligned reads at a location. By measuring uniqueness across a genome, a mappability track can be generated.

Mappability tracks are commonly used to normalise depth of coverage, a particularly important pre-processing step for accurate prediction of duplication events (Lai & Ha, 2014). Regions falling below a mappability threshold are often removed from subsequent analyses to reduce the incidence of false positives (Cheung et al., 2011). This strategy can exclude a significant proportion of the genome:

approximately 20% of the human genome is considered inaccessible due to poorly mapped regions (The 1000 Genomes Project Consortium et al., 2010). Repeated regions also exhibit high levels of variability (Medvedev et al., 2009): excluding these regions necessarily prevents many real variations from being discovered.

More specific filtering approaches, particularly with respect to ASE, have been undertaken. Another approach is to simply not consider any site containing too much variation. Stevenson et al. (2013) eliminated bias by excluding any site with more differences than the maximum tolerance of the aligner. This approach breaks down if the donor diverges too much from the reference: an increasing proportion of the genome is excluded from the analysis.

All of these filtering approaches have the disadvantage of eliminating a potentially significant number of valid variations.

3.2.3 Alignment Strategies

Due to the computational complexity of the approximate string matching problem, aligners exploit shortcuts and heuristics. Aligners include configurable parameters which may reduce reference bias. Stevenson et al. (2013) generated simulated reads and then artificially introduced known SNPs into the reads. Reference bias was calculated by comparing the number of mutated reads that successfully aligned against the number of unchanged reads that successfully aligned. Increasing the maximum number of allowed mismatches significantly reduced reference bias in this context, but this solution applies only to known SNPs.

A comprehensive study comparing a range of aligners with numerous configuration options and experimental conditions found that, in general, read length and the *choice of reference* have more impact on alignment accuracy than the choice of alignment software, or the choice of alignment parameters (Hatem et al., 2013). Nevertheless, hash-based aligners tend to be more sensitive than suffix-array based aligners and subsequently are preferred for divergent sequences (Gontarz et al., 2012).

A limitation of alignment is that each read is considered *independently*. Consequently, each aligned read may suggest a variation that is inconsistent with other overlapping reads. This commonly occurs around indels. Individual reads often suggest SNVs that conflict with other reads, rather than suggesting a set of compatible alignments. Local realignment considers overlapping reads together. Applying this to likely indel sites helps recover indels while reducing false positive SNV calls (Albers et al., 2011; DePristo et al., 2011).

Similarly, novel insertions that are longer than the read length cannot be correctly aligned to the genome and can only be recovered with paired-end reads or by employing de novo assembly techniques (Alkan et al., 2011). Although de novo assembly suffers from issues such as high sensitivity to errors and repeats, hybrid solutions combining both assembly and alignment are common (Schneeberger et al., 2011; Bao et al., 2014; Holtgrewe et al., 2015).

Modern alignment tools report a degree of confidence in the correctness of each aligned read.

Confidence (or *mapping quality*) is a reflection of the existence of alternative feasible locations for the read on the reference genome (Li et al., 2008). A read that aligns equally well to multiple locations will report low mapping quality, while a read with a single high-scoring alignment will report high mapping quality. Although a high-scoring alignment implies few mismatches between the read and the reference sequence, this does not necessarily translate to high mapping quality: a read that aligns equally well to multiple locations will report a low mapping quality, regardless of the alignment score.

When assessing a set of aligned reads, mapping quality enables each read to be appropriately considered with respect to the aligner's confidence in its correctness.

3.2.4 Reference Augmentation

Compared with filtering, augmentation modifies problematic regions rather than excludes them. Several approaches that modify the reference genome to reduce bias and increase accuracy have been proposed.

Degner et al. (2009) modified the reference by masking all known variants. Given a reference allele and an alternative allele, each variant location was replaced with a third unrelated allele, thus ensuring that neither the reference nor the alternative variant could be preferred during alignment. This approach successfully removed bias toward the reference at these sites, however, this did not reduce the number of strongly biased alleles. The extra guaranteed mismatch resulted in more incorrectly aligned reads and introduced other biases.

Related to this approach of ensuring equal representation is the construction of a genome that includes all possible variations. There have been several attempts to incorporate known variations into the reference genome. At each site of interest, Satya et al. (2012) appended the alternative variation to the reference genome, thus enabling reads to align equally to either possibility. Similarly, *BWB-BLE* (Huang et al., 2013a) combines these approaches by enabling SNPs to be represented as part of the reference sequence with IUPAC codes (Cornish-Bowden, 1985), and by appending other known variations to the end of the reference genome.

These methods suffer from important limitations. In order to enable a read to align completely to an augmented subsequence, each variant is padded on each side with sequence equal to the read length. The number of known SNPs exceeds 140 million, and the number of known indels exceeds 16 million. The number of known variants continues to grow; current read lengths are typically 100 bp and are also expected to increase dramatically. Additionally, variants tend to cluster, resulting in multiple variants appearing on a single read. Augmenting the genome with every possible combination of variants to allow unbiased alignment is unscalable.

Rather than attempting to package all known variants into a genome, iterative solutions generate a pseudogenome incrementally by repeatedly performing an alignment and then incoporating discovered variants into the reference genome (Ghanayim & Geiger, 2013; Huang et al., 2013b). Iterative approaches are common when performing multiple genome alignment (Darling et al., 2010), but the additional computational requirements and expected improvement to read alignment have not been

deeply studied. Ghanayim & Geiger (2013) found improvements iterating *BWA* with sequence divergence between 4% and 7%, but no comparison was made to more sensitive aligners that are suited to divergent sequences.

A final method of augmentation adds sequence intended to attract reads that are expected to align incorrectly. In many assemblies, there are sequences that are known to exist somewhere in the genome but have not yet been integrated into the reference sequence. In order to prevent reads from these regions from mapping erroneously to the reference and generating artifacts and false variations, the reference sequence is augmented with decoy sequences to collect these reads (Genovese et al., 2013). Recent versions of the human reference genome include alternative sequences for regions that exhibit high levels of variability. This helps mitigate the difficulty of representing these regions with a single sequence (Rosenbloom et al., 2015).

3.2.5 Multiple Genome Solutions

Conventional aligners map reads to a single reference genome. This limits which variations can be discovered, since it is difficult or impossible to confidently predict variations in regions that diverge from the reference. Rather than augmenting an existing single reference genome, a potentially more powerful solution is to enable reads to be aligned to a collection of reference genomes, enabling reads to be aligned to an integrated multi-genome. By representing more of the donor genome in a collection of references, the incidence of highly divergent sequence in the donor genome is reduced, thus enabling more accurate alignment and ultimately, more accurate reconstruction of the donor genome.

GenomeMapper (Schneeberger et al., 2009) and *GCSA* (Siren et al., 2011) use compressed data structures that enable the indexing and storage of a collection of genomes, so that reads can be mapped to any known variation that is present in any of the available genomes. Although this approach potentially enables the integration of all kinds of variations present in the collection of genomes, these schemes have excessive memory requirements for longer sequences such as the human genome. *GCSA* requires in excess of 1 Tb of memory to index the human genome, while *GenomeMapper* requires 19 Gb per genome.

A graph-based data structure is an intuitive model for representing multiple related genomes, but generating a multi-genome representation is an NP-hard problem (Dilthey et al., 2015). Despite this, there has been considerable progress towards this goal. Kehr et al. (2014) demonstrate a representation based on Cactus graphs, but a practical solution enabling read alignment to multiple genomes is not yet available.

3.2.6 Summary

Each of these approaches to reducing reference bias has limitations. Statistical methods are limited by the information that is available in an alignment and cannot recover information that has already been lost due to bias. Filtering can eliminate bias, but at the expense of excluding difficult

regions. Fine-tuning alignment parameters is limited because short reads necessarily create ambiguity in genomes containing repeated sequence, regardless of the alignment strategy. Augmenting the reference sequence is limited by the extent to which multiple sequences can be incorporated into a single sequence. Finally, multi-genome solutions offer the most flexible approach, but a practical solution is not yet available.

3.3 Sources of Reference Bias

Although systematic bias is recognised as an issue and solutions have been proposed to reduce it, there has been limited investigation of the causes of reference bias. An alignment task consists of four components that contribute to reference bias:

- *Reference genome*: genome structure can have a significant impact on reference bias. Areas of repeated sequence or low complexity (or both) are typically underrepresented or misrepresented in read counts. The effect of genome structure on alignment is referred to as *mappability* or *mapping bias*.
- *Mapping strategy*: alignment tools make assumptions and apply heuristics when selecting the most likely position for a read. This results in systematic biases.
- *Donor genome*: a critical factor affecting alignment accuracy is the size, complexity, and distribution of variations separating the donor genome from the reference genome.
- *Read generation*: the sequencing process generates various biases and other artifacts that affect alignment accuracy. Sequencing errors effectively increase the edit distance between the reference genome and the donor genome.

Each of these components are related. For example, if sequencing generates predominantly insertion errors, then an alignment tool that supports gapped alignment is likely to be less biased.

3.3.1 Reference Genome

The process of alignment involves placing fragments from a donor genome accurately on a reference genome. The structure of the reference genome influences the feasibility of this task.

Consider a simplified example. Error-free reads of length k are taken from a genome, and those reads are then aligned back to that same genome. Consequently, there are no errors and no variation. This scenario is illustrated in Figure 3.1. Is it possible to accurately align reads to the reference and, ultimately, reconstruct the original genome?

In the absence of differences in the form of errors or divergence from the reference, a read will map correctly and unambiguously to its original location if there is no other identical substring on

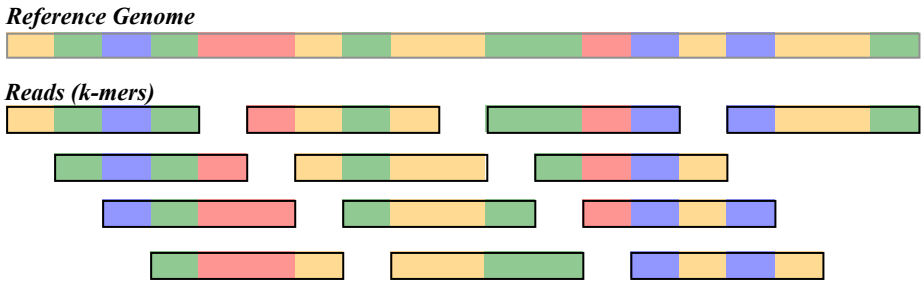

Figure 3.1: Alignment: when every possible k-mer uniquely matches a location on the genome there is no ambiguity.

the genome. An added complication is that DNA consists of double stranded, complementary sequences, and reads can originate from either strand. Subsequently, for every possible read to map unambiguously to its correct location, every possible k-mer from both strands must be unique.

More generally, if a read sequence exists in multiple locations on the genome, either as the same sequence or its reverse complement, then without further information the read cannot be aligned unambiguously, and is referred to as a *multi-read* (Treangen & Salzberg, 2012).

The frequency and length of repeated sequences in the reference genome is a critical factor determining the proportion of reads that can be confidently aligned. A reference genome consisting entirely of the same repeated base is ambiguous at all locations and all read lengths: it provides no extra information about the relative or absolute position of reads. In contrast, a reference with no repeated k-mers enables all reads to be confidently aligned and enables the donor genome to be reconstructed with no uncertainty. Real genomes fall between these two extremes.

A method of reducing the incidence of ambiguous reads is to increase the read length. Only reads that are completely contained by repeated sequence are ambiguous: if a read is long enough to span all repeated sequences then ambiguity is eliminated. As a result, various analyses have calculated levels of repetition across several genomes for different read lengths. Long repeats of length 500 bp to 8000 bp cover 21% of the human genome (Treangen & Salzberg, 2012). Li et al. (2014) considered k-mers in the range of 20 bp to 1000 bp, finding that even for a read length of 1000 there is significant ambiguity. Plant genomes contain an even greater proportion of repeats (Schnable et al., 2009).

3.3.2 Mapping Strategy

In addition to ambiguity due to duplicated regions, reads are frequently incorrectly aligned due to sequencing errors or variations at repeated regions (Li et al., 2008), a problem compounded by the relatively high incidence of variations in these regions (Sharp et al., 2005).

Aligners apply different strategies to multi-reads, each of which has important implications for bias and variation discovery. Common strategies include:

- *Discard all multi-reads:* this limits the analysis to unique regions of the target genome, which

reduces specificity, since variations in non-unique regions cannot be detected. This strategy can also generate more false positives: Abyzov et al. (2011) found that discarded reads in repetitive regions caused depth of coverage analysis to incorrectly predict many false deletions.

- *Align the read to all potential positions:* this strategy over-reports read depth, resulting in fewer false positives but also fewer true deletions.
- *Randomly choose from one of the locations:* This is the default behaviour of *Bowtie 2* (Langmead & Salzberg, 2012) and *BWA-MEM* (Li, 2013). Rather than systematically over- or under-reporting coverage, this strategy effectively dilutes the signal indicating a duplication or deletion. Both deletions and duplications are lost by the random distribution of reads (Teo et al., 2012).

When a read does not exactly match any part of the reference, approximately matching loci are considered. A consequence of using a score-based model is that the highest scoring alignments tend to be biased towards those that are considered most likely (Lunter et al., 2008), even though less likely configurations do actually occur, albeit at lower frequencies. Artifacts of this approach include gap wander, gap attraction, and gap annihilation (Holmes & Durbin, 1998), and are a result of treating each read independently when aligning to a single reference. Local realignment only tends to reduce the incidence of false positive insertions (Homer & Nelson, 2010).

If a read varies too much from the reference, the aligner will either not map the read at all, or clip regions containing too much variation. This biases the set of aligned reads against regions containing high levels of variation. In particular, some reads that correctly align to the reference contain enough variation such that any additional variation will prevent the read from being aligned. This introduces bias that reflects the maximum threshold of variation that the aligner tolerates. Increasing the aligner's tolerance for mismatches and excluding areas of high variation can reduce this type of bias (see Section 3.2.4).

The behaviour of aligners varies considerably with respect to allowed mismatches and penalties for different type of variation (see Table 2.1). These settings bias the alignment results towards particular types of variation. For example, when comparing *Bowtie 2* and *BWA-MEM*, *Bowtie 2* is less affected by CNVs but more susceptible to false negatives when calling SNVs (Li, 2014).

Finally, a fundamental limitation of current aligners is their consideration of each read independently. This can result in individual reads being optimally aligned to the reference that are clearly incorrect when considered in the context of other mapped reads in the region. This effect is particularly prevalent when indels are present in repetitive regions; these variations are often aligned as discordant multiple mismatches (Narzisi & Schatz, 2015).

3.3.3 Donor Genome

The set of differences between the donor genome and the reference genome is an important consideration. Individuals of the same species are often extremely similar: unrelated humans average 99.8%

similarity (Reinert et al., 2015). However, variations are distributed unevenly and are frequently clustered (Montgomery et al., 2013).

If a region varies significantly from the reference, the aligner may fail to correctly map reads spanning this region. Ambiguity arises: a region of low coverage may indicate an area of high variation, but it can also indicate a deleted region, or a repeated region. Low coverage suggests that the aligner cannot find reads that satisfy the requirement of being both sufficiently similar to this region, and sufficiently dissimilar to the rest of the genome.

The type of variation is also significant. Modern aligners tolerate a small number of substitutions and indels. Short variations are directly indicated by the alignment, but longer rearrangements that extend beyond the length of a single read are also common. Longer rearrangements must be inferred by interpreting clusters of aligned reads, a task made more challenging as the distance from the reference increases.

3.3.4 Read Generation

Although the process of generating reads introduces biases, this study only considers the impact of *sequencing errors*. Sequencing errors are effectively noise that increases the distance between the read and the correct location on the genome, while potentially decreasing the distance to other candidate positions. We assume errors are uncorrelated and unbiased, and consider the most common types of sequencing errors: single base changes and indels.

3.4 Types of Reference Bias

Alignment is effectively a task of assessing similarity. A short sequence is placed at what is considered to be the most similar region in a larger sequence, or, if no similar region exists, the read will not be aligned. By considering this process and the factors influencing this process, we can categorise the types of bias (see Figure 3.2):

- *Similar reference regions*: if the correct location for a read is mirrored by other identical regions on the reference, the aligner cannot distinguish the correct location from other incorrect locations. Similarly, if the correct location is mirrored by other very similar regions, the alignment becomes sensitive to errors and variations, since small changes to the read can result in the read matching another similar region.

- *Mutation to similar region*: a read will be correctly aligned if the difference between the read and the correct location's subsequence is less than the difference between the read and any other potential location. A variation or error present in a read increases the difference between the read and the correct location, which increases the likelihood that another position will appear more similar than the correct location.

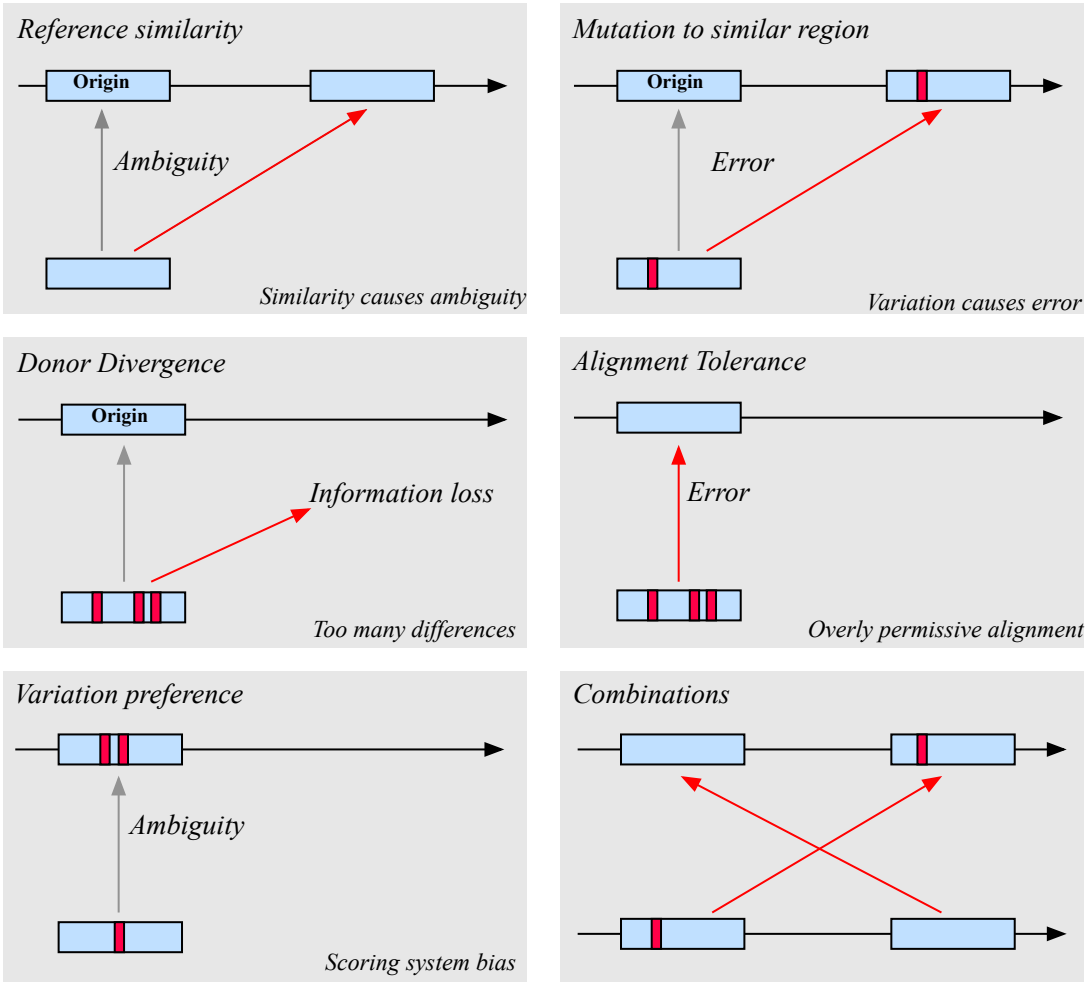

Figure 3.2: Types of bias. Bias arises from sequence similarity, sequence divergence, and alignment assumptions. Red marks indicate differences between the two genomes.

- *Divergence*: if a read contains significant levels of variation, it may exceed the tolerance of the aligner, resulting in an unmapped read. This results in lost information, with areas of high variance being the most affected – which are also usually of greatest interest. If aligners have a tendency not to align highly divergent reads, the alignment will be biased towards the reference. For example, long novel insertions result in reads that do not contain any sequence from the reference, and cannot be aligned to the reference. The result is the loss of novel content.

- *Alignment tolerance*: in contrast to the problem of losing divergent sequence, if an aligner is too tolerant of differences, sequences that should not align to the reference may be incorrectly aligned. Sequence data often contains contamination and other content that has no correct location on the reference genome. Incorrect alignment occurs if the distance from one of these reads to a location on the reference is less than the maximum tolerance of the aligner.

- *Similar donor regions*: similar to repeated sequence on the reference, repeated sequence on the donor results in ambiguity. Inferring the true location of repeated sequence is impossible without further information. As with similar reference regions, similar donor regions increase the sensitivity of the alignment to variations and errors.

- *Variation preference*: aligners use scoring systems to calculate distances between subsequences to assess similarity. Scoring systems introduce bias because particular types of variation are implicitly favoured. For example, point mutations are usually penalised less than indels, resulting in less evidence being required for a point mutation to be predicted.

- *Combinations*: each of these biases can interact, which increases the difficulty of detection.

Alignment parameters often affect each category differently, leading to trade-offs between each type of bias. For example, bias due to divergence can be reduced by increasing the tolerance of the aligner, however, this increases the bias that arises due to contamination and errors.

3.5 Measurement of Reference Bias

The goal of alignment is to place reads from a donor genome correctly on to the reference genome. For each read, this process has the following possible outcomes:

- *The read is correctly aligned to the reference.* Assuming that a correct alignment for the read exists, this outcome increases evidence for the correct interpretation of the donor genome. However, not all reads have a corresponding location on the reference. Novel sequence has no corresponding sequence on the reference.

- *The read is not aligned to the reference.* If the read originates from the donor genome and has a corresponding location on the reference genome, then any evidence this read may have provided about its location is lost. In contrast, this is the correct behaviour if the read does not

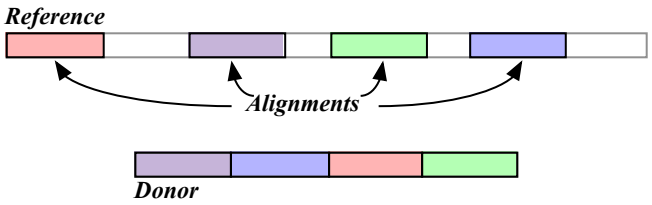

Figure 3.3: Accurate alignment does not imply that the donor genome can be reconstructed. Gaps prevent the relative order of reads from being inferred.

originate from the donor, is a sequencing artifact, or originates from part of the donor that is not represented on the reference.

- *The read is incorrectly aligned to the reference.* Incorrectly aligned reads add noise to the resultant alignment, potentially masking evidence suggesting the correct sequence. They can also generate false artifacts and suggest alternative, incorrect sequences.

The accuracy of an alignment can be measured in terms of these three outcomes and is a common approach in comparisons of sequence alignment software (Mäkinen & Rahkola, 2013; Ruffalo et al., 2011; Schbath et al., 2012). A read is classified based on whether it aligns to the correct location (true positive), aligns to an incorrect location (false positive), is unmapped but does not originate from the donor genome (true negative), or is unmapped and does originate from the donor genome (false negative). This enables the accuracy of an alignment to be assessed using measures such as sensitivity and specificity. However, even if all reads align perfectly to the reference genome, successful recovery of the donor genome does not necessarily follow. If the donor is significantly shorter than the reference then the donor sequence is unlikely to be recoverable, even with accurate alignment (see Figure 3.3). In contrast, if the donor is significantly longer than the reference then the donor either contains novel content relative to the reference, or repeated sequence. Novel content is unlikely to align correctly, while the locations of repeated sequences are often difficult to recover. Consequently, many studies assume only small-scale variations between donor and reference (for example, Hatem et al. (2013); Ruffalo et al. (2011); Schbath et al. (2012)).

A better metric is to measure how accurately the aligned reads reflect the differences between the genomes. The alignment (or non-alignment) of each read provides evidence suggesting similarities or differences between the donor and the reference. Taken together, the set of aligned reads suggests a set of differences between the genomes of interest. Differences can be localised (point mutations and indels) or large-scale rearrangements (SVs). Regardless of the type of difference, reconstruction of the donor genome from a set of differences is a straightforward process (Huang et al., 2013b). Hence, if the set of differences is known, then the donor genome can be reconstructed. We can measure our ability to reconstruct the donor genome by measuring the accuracy of the set of differences suggested by the alignment. Consequently, to evaluate our ability to accurately infer the donor genome from a set

of reads, we classify each read based on whether it correctly indicates the existence or non-existence of each variation.

When interpreting the evidence generated by aligned reads, we consider reference bias to be the *false negative rate*: the existence of a variation that is not indicated by the evidence. However, minimising the false negative rate by reducing the required level of evidence introduces false positives: a common classification trade-off. The choice of threshold separating true variations from noise reflects the trade-off between limiting the number of missed variations and limiting the number of incorrectly predicted variations.

Variation discovery algorithms that assess the evidence generated by alignment are an active area of research (Bartenhagen & Dugas, 2015; Escaramís et al., 2015). It is difficult to estimate how much evidence is sufficient for a variation to be accurately discovered by any algorithm, but even a perfect variation inference algorithm cannot discover variations that are not supported by any evidence at all. Consequently, of particular interest in this study is the class of variations for which supporting evidence is absent in the set of aligned reads.

3.6 Summary

Reference bias is a significant issue that arises in numerous circumstances. Although solutions to combat reference bias exist, none are completely satisfactory. No existing solution eliminates the problem. Reference bias arises as a result of the structures of the genomes of interest, limitations of alignment, and sequencing errors.

Accurately reconstructing the donor genome is equivalent to the problem of determining the differences between the donor and the reference. Consequently, we focus on evidence suggesting differences between the genomes. We define reference bias to be the false negative rate when evaluating evidence for a variation. Of particular interest are variations with no supporting evidence. Reference bias occurs when a variation exists in the donor genome that is not indicated by any evidence in the set of aligned reads.

In the following chapter, we consider the contribution of the structure of the reference genome to reference bias, and identify scenarios where genome structure significantly influences alignment outcomes.

Chapter 4

Genome Structure

We have defined reference bias and identified its causes. The first contributor to reference bias that we consider is the effect of the structure of the reference genome.

4.1 Method

In this simplified scenario, we consider the donor genome to be identical to the reference genome: there are no variations between the genomes. The only impediment to alignment is the content of the reference, and any noise (read errors) that may arise in the sequencing process.

If all possible k-mers are unique, then there is no ambiguity in the alignment and any read from any location can be aligned uniquely (see Section 3.3.1). In contrast, if the genome is a sequence of identical bases (a homopolymer), then all locations are indistinguishable. A naïve approach might be to assume that genomes can be modelled by a random distribution of bases, but genomes are far from random. The expected longest repeated subsequence of a DNA string of length n with randomly distributed bases is $log_4 n$ (Mäkinen et al., 2009). For example, *E. coli* K-12 MG1655 has a length of 4 641 652 bp and its longest repeated substring is 2815 bp, but the expected longest repeated substring of a string of this length is just 11 bp.

Any repeated substring longer than the read length potentially creates ambiguity, since shorter reads falling in repeated regions map equally well to multiple locations. If an unpaired read exactly matches multiple locations on the genome, each location is equally likely. Alignment software assigns a mapping quality to each aligned read to indicate its confidence in the correctness of the location. To test the effect of similar sequence across the reference genome, and the effect of non-matching reads, we inserted repeated sequence with a variable number of mutations at multiple locations on the *E. coli* K-12 MG1655 reference genome, then artificially generated 100 bp single-ended reads, and varied the following parameters:

- *Difference between the correct location and an alternate almost identical location.* We applied a variable number of point mutations to the alternate location.

- *Difference between the correct location and the read.* We applied a variable number of point mutations to the read.
- *Content of the inserted sequence.* We generated low complexity content by inserting homopolymer runs, STRs, and by varying the level of GC-content.
- *Position of the difference on the read.* We measured the impact of the position of the mutation in the read, and the position of the repeat within the genome.

We aligned the simulated reads to the modified reference genome with *Bowtie 2* and *BWA-MEM*. For these experiments, we used homozygous point mutations to represent differences between sequences. Point mutations are the simplest, most common, and most studied type of variation.

Aligners report mapping quality with different scales: the maximum reported by *BWA-MEM* is 60, while the maximum reported by *Bowtie 2* is 44. We normalised mapping quality based on the highest mapping quality reported by each alignment tool.

To evaluate the impact of structure on alignment, we profiled the incidence of repeated sequence on several publicly available reference genomes (see Appendix A) with varying characteristics. We included representatives from viruses (*Human immunodeficiency virus 2*), bacteria (*Escherichia coli*), and eukarya (*Homo sapiens*). Since genome content varies widely between species, we also included *Plasmodium falciparum* due to its extreme GC-content (Gardner et al., 2002), and *Drosophila melanogaster* for its high levels of repeated sequence (Adams et al., 2000).

In particular, we measured the number of potential reads originating from the genome that have ambiguous locations. This provides an indication of how rapidly ambiguity increases if alignment software tolerates more mismatches.

4.2 Results

Figure 4.1 illustrates the relationship between the number of read mutations and the number of mutations in the alternative location. Two main factors affect the reported confidence of an aligned read:

- The similarity between the correct location and other locations on the genome.
- The difference between the read and the correct location.

Confidence increases as the edit distance from the correct location to the most similar location increases, but confidence is still significantly affected by alternatives with several mutations. Mapping quality is significantly reduced by alternative subsequences with less than 5 differences with *BWA-MEM*) and less than 10 differences with *Bowtie 2*.

Confidence is also significantly affected by differences between the read and the correct location. In particular, when a similar location exists on the genome (within 5 differences), the mapping quality of a read with a single mutation is typically half that of an exactly matching read. Reads begin to fail

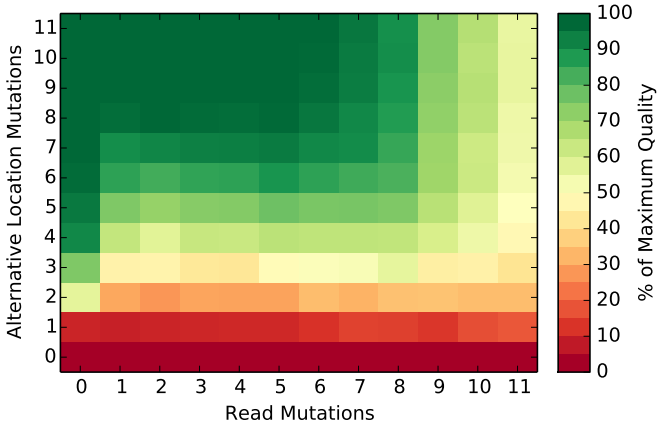

Figure 4.1: BWA's reported mapping quality as a function of differences between the read and differences between an alternative location. Similar alternatives guarantee low mapping quality, but if the read has many mutations then more distant alternatives also affect mapping quality. Mapping quality has been normalised to the maximum reportable value.

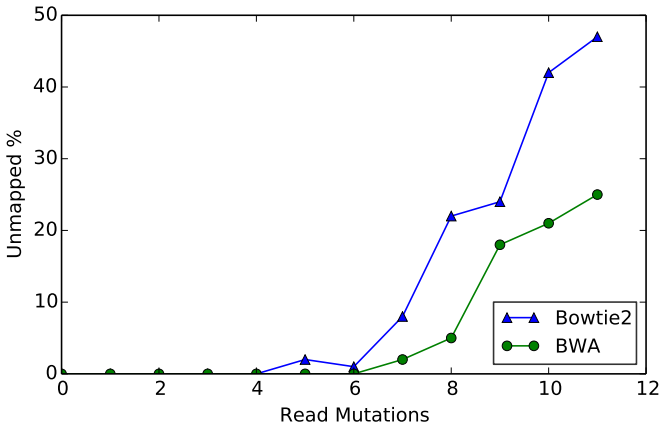

Figure 4.2: Effect of read errors on proportion of unmapped reads. Unmapped reads are negligible when the read contains less than 5 mismatches but rapidly increase with 7 or more mismatches.

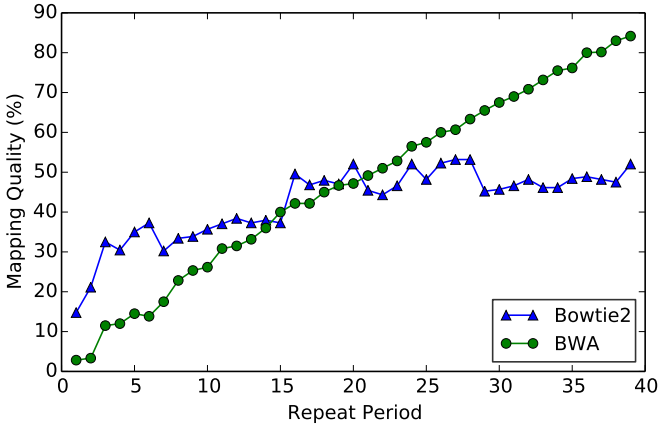

Figure 4.3: Effect of STRs on reported mapping quality (normalised). Short repeat periods have the greatest impact on mapping quality. BWA-MEM *is particularly sensitive to short STRs.*

to be aligned once they contain 5 or more mismatches with *Bowtie 2*, and 7 or more mismatches with *BWA-MEM* (see Figure 4.2). The level of unmapped reads is independent of the presence of repeated sequence and is purely dependent on the difference between the read and the nearest feasible match on the genome. We explore this in further detail in Section 5.2.1.

To a lesser extent, the underlying complexity of the sequence content affects the confidence in the alignment. Extreme levels of GC-content (either high or low) only affect alignment if the entire read is a homopolymer. In contrast, short repeated sequences covering the read are penalised, particularly STRs (see Figure 4.3). However, this effect is weaker in *Bowtie 2* and in more recent versions of *BWA-MEM*. The location of the mutation within the read has minimal impact on alignment, but any mutation within 4 bp of the end of a read is clipped by *Bowtie 2*. We explore the impact of clipping on variation prediction in Section 5.2.3.

4.2.1 Application to Real Genomes

Exact repeats that are longer than the read length generate obvious ambiguity, but near-repeats are also ambiguous. Minor variations in the donor genome or errors in the sequencing process can result in the alternate location appearing more likely. Reads that map to near-repeats are given low mapping quality scores due to ambiguity.

We calculated the proportion of reads affected by exact repeats and near-repeats on existing reference genomes by generating reads from all possible loci and then comparing all possible reads to every other possible read. Table 4.1 illustrates that the number of reads affected by near-repeats is significant, suggesting that as well as exact repeats, near-repeats are also common. Pseudogenes and transposable elements are particularly prevalent in eukaryotic genomes (Kazazian, 2004). This gives rise to significant levels of near-repeats.

Genome	Exact Match	1 mismatch	2 mismatches	3 mismatches
D. melanogaster chr2L	15.3%	17.6%	20.9%	23.2%
E. coli K-12 MG1655	2.3%	2.5%	2.8%	2.9%
H. sapiens chr6	1.0%	2.2%	6.3%	14.2%
H. sapiens chr21	0.5%	1.1%	3.6%	7.9%
P. falciparum	4.1%	4.7%	7.0%	8.9%

Table 4.1: Percentage of reads affected by near-repeats with a read length of 100 bp. We generated all possible 100 bp reads for the specified genome (or chromosome), then calculated the proportion that were very similar (within 3 mismatches).

4.3 Summary

Although the effect of repeats has been studied previously (see Section 3.3.1), we also quantified the impact of near-repeats. Near-repeats significantly affect the confidence assigned to an aligned read. We found that the prevalence of near-repeats varies greatly across and within species, and that the incidence of exact repeats is not a good predictor for the incidence of near-repeats.

Reads that differ substantially from the reference tend to be unmapped, or mapped with low confidence. With the two tested tools, reads with up to 4 mismatches are usually correctly mapped, but the proportion of unmapped reads increases rapidly once a read contains 7 or more mismatches. An assumption of the alignment process is that the donor genome does not vary significantly from the reference genome. In the next chapter, the limits of this assumption are investigated by synthetically introducing variations into the donor genome.

Chapter 5

Short Variations

We have investigated the limitations of alignment due to the structure of the reference genome. We next consider bias that arises due to differences between the reference and the donor – specifically, short variations. Alignment assumes that most reads map correctly, an assumption that is likely to break down in regions where the donor diverges from the reference.

In this chapter, we assess the ability of aligners to map reads containing short variations, with the objective of determining the likelihood of correctly predicting a variation given the output from the alignment software.

5.1 Method

In order to evaluate the accuracy of alignment, a set of known variations between two genomes is required, but obtaining a true set of naturally occurring variations is challenging. For example, cancer genomes consist of numerous mutations relative to the unaffected DNA, but the true set of differences is not known. Consequently, we employ simulation to investigate the factors affecting read mapping accuracy and variation detection. This enables us to measure the ability of alignment software to align specific variations.

Figure 5.1 summarises the simulation pipeline, which performs the following operations:

- We add repeated sequence to the genome so that the relationship between repetition and short variation prediction can be determined.
- We then synthetically apply short variations to form a donor genome. We investigate two models of variation distribution: randomly distributed and uniformly distributed.
- Simulated reads are generated from this donor genome. We tag each read with its original location, and any variations that it spans. This establishes a ground truth. We vary read length from 100 bp to 2000 bp, which includes the typical fragment length for paired-end reads (300 bp to 500 bp). Coverage is varied from 10x to 100x with both uniform and Poisson distributions.

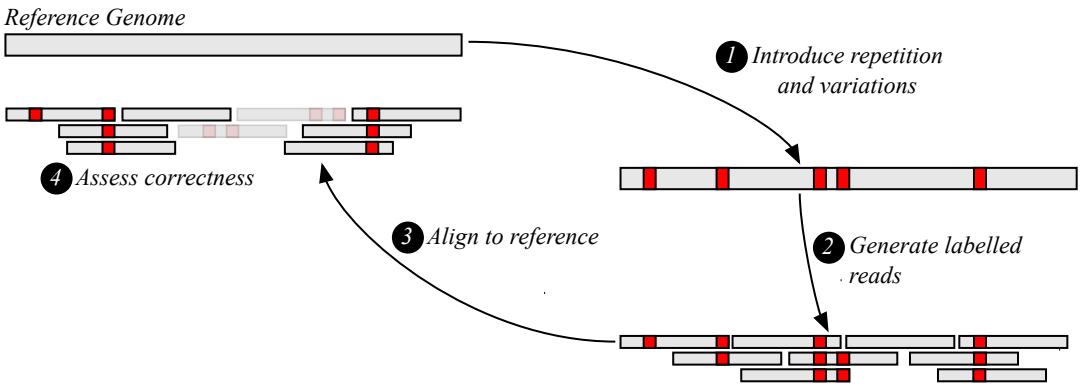

Figure 5.1: Alignment evaluation pipeline. A reference sequence is mutated with variations, represented here as red marks. We generate simulated reads from the mutated genome and align them to the reference, then evaluate the output.

- We align the reads to the reference genome with *BWA-MEM* and *Bowtie 2*. The reference genome for these experiments is *E. coli* K-12 MG1655.
- The accuracy of the alignment process is assessed by comparing the aligned position to the original correct position.

For this analysis, we assume single-ended reads and homozygous mutations. Unless otherwise specified, mutations and errors are randomly distributed. Coverage variation is an artifact of the sequencing process (Ross et al., 2013). Consequently, we run simulations with both a uniform distribution of reads, and a Poisson distribution of reads. Uniform coverage eliminates all bias, while a Poisson distribution is a more realistic unbiased approximation of generated reads. Lander & Waterman (1988) demonstrate the relationship between these two distributions: if read sampling is approximately Poisson then sufficient sequencing can guarantee any minimum uniform coverage depth.

The parameters of this simulated experiment are generally favourable compared to the task of inferring naturally occurring variation between genomes. Heterozygous mutations are more difficult to accurately predict than homozygous variations. We simulate randomly distributed mutations and errors, but actual mutations and errors tend to cluster, which leads to regions of relatively high variability. Since paired reads can resolve repeats, we simulate read lengths that span typical DNA fragment sizes. A read that spans a DNA fragment provides at least as much information as a paired-end read (see Section 2.2). Consequently, the results presented here provide an optimistic upper bound on variation prediction accuracy.

Using this simulation framework, we vary the mutation rate, error rate, read length, and coverage.

5.1.1 Inference of SNVs

Although accurate alignment is an important requirement for imputation of the donor genome, a more appropriate metric is the ability to accurately infer differences between donor and reference from the

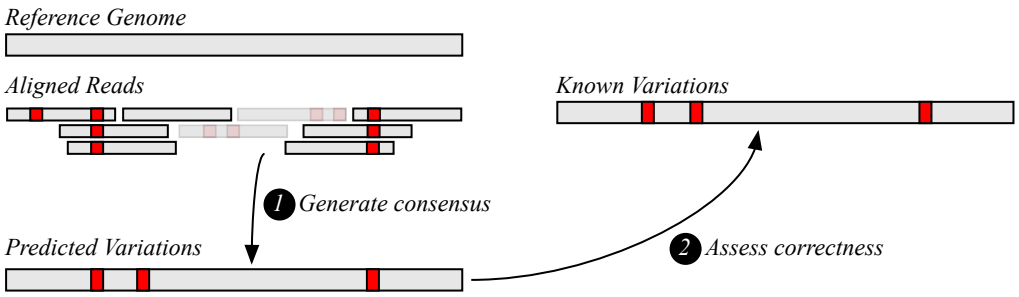

Figure 5.2: Variation evaluation pipeline. Variations are imputed from the mapped reads and compared to the known variations.

set of mapped reads (see Section 3.5). By comparing the set of predicted variations to the set of true variations, we can measure how accurately the donor genome can be recovered. This enables us to identify instances where this process fails.

Figure 5.2 illustrates this process. It consists of the following steps:

- *Find variations*: The imputation of SNVs from aligned data is based on the evaluation of evidence at each base on the reference genome. A pileup of overlapping reads at each genome position is analysed to determine whether a difference between the donor and the reference exists (see Figure 2.5). If the majority of bases at a location vary from the reference, a variation is recorded. The confidence for the call is calculated by measuring the proportion of all predicted bases represented by the majority base, with simple additive smoothing (Chen & Goodman, 1996) applied to smooth confidence calculations in low coverage regions. If no read suggests the existence of a variation, then it is considered to be lost, regardless of the sensitivity of the variation discovery process.

- *Compare variations*: Given the set of true variations and the set of predicted variations, true positives are the intersection of the two sets, false positives appear only in the predicted variations, while false negatives appear only in the true variations. We focus on false negatives: we assume that post-processing and algorithmic improvements can find any variation for which there is some evidence in the alignment.

5.1.2 Inference of Indels

As with SNVs, indels are likely to be fully contained in several reads, thus enabling their existence to be directly inferred from the alignment results. This is a common approach taken by tools such as *SAMtools* (Li et al., 2009a) and *Dindel* (Albers et al., 2011) to detect indels. Longer indels are considered to be SVs and, since they are not contained within a read, require the interpretation of other types of evidence (Schröder et al., 2014).

Detection of indels introduces additional difficulties compared to the task of identifying SNVs. The presence of indels in a sequence necessarily introduces gaps into the alignment. As a result, prediction can be sensitive to alignment parameters and is prone to reduced accuracy in areas of low-coverage or low-complexity (Rizk et al., 2014). Lunter et al. (2008) demonstrated that alignment biases in the mapping process are a significant issue in this context, regardless of the alignment parameters.

A further issue is that, unlike SNVs, indels cannot always be uniquely identified by position. For example, if one repeated segment amongst a series of repeats is deleted, any one of the repeats could be identified by the aligner as the deleted fragment. To mitigate this issue, we allow the indel to appear anywhere in the read.

We only consider direct evidence for a variation. The output from an alignment includes a CIGAR field for each read, which explicitly describes predicted deletions and insertions for that read (Li et al., 2009a). Each indel prediction is assigned to the corresponding position on the reference genome. If the majority of reads support that indel prediction then the variation is called with a confidence based on the proportion of supporting reads.

We assess the ability of aligners to map reads containing indels, with the goal of measuring how much evidence for an indel is indicated by an alignment, and the factors that influence that evidence. In addition to evaluating mutation rate, error rate, read length and coverage, we also assess the impact of indel length and surrounding content on the accuracy of indel prediction.

5.1.3 Mappability

Section 3.2.2 introduced the concept of mappability, which measures the relative expected depth of coverage of reads at any given location in the genome. Mappability is commonly used to normalise biases in coverage depth that arise due to the structure of the genome. Although regions of repetitive content or low complexity produce low mappability scores, mappability is not necessarily an accurate indication of variation prediction accuracy.

We extend the concept of mappability by introducing variations into the process, enabling us to characterise regions of the genome that have significant impact on variation prediction ability. In particular, by looking at individual variations, biases specific to each type of variation can be identified.

For every location on the genome, consider the set of all reads that span that location. For a location l and read length r, excluding locations within r bases of the start of the genome, this will be the set of reads originating from $l - r + 1$ to l. By artificially generating all possible reads that span a location and aligning these to the reference, an indication of the mappability of the location can be determined. The *mappability* of a location can then be defined as the proportion of all possible reads spanning that location that when aligned, map to their correct location.

Chapter 4 considered biases that arise due to the structure of the reference genome, and these biases are directly reflected by existing mappability measures. However, these measures do not reflect

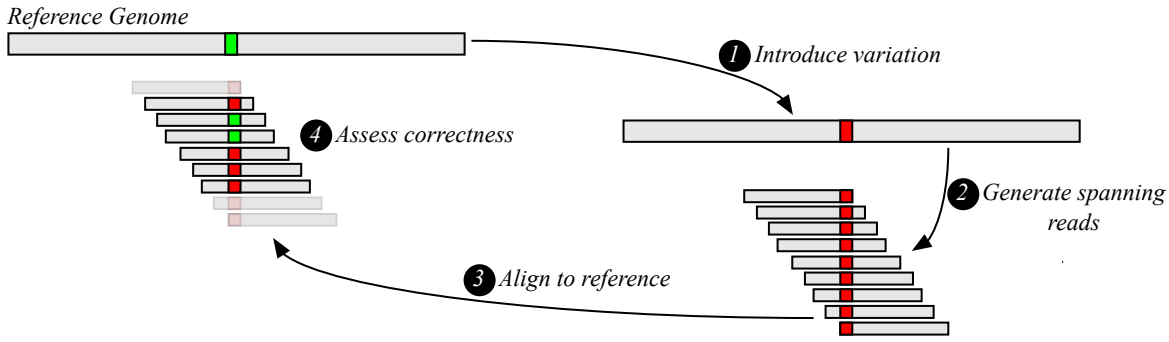

Figure 5.3: Process for measuring the mappability of a specific variation across a genome. We generate every read spanning a variation, then measure the proportion of reads that suggest the variation.

the ability of specific variations to be accurately inferred. In particular, regions of high mappability do not necessarily translate into regions where variations can be confidently predicted.

To measure mappability relative to a variation, at each location, the variation is synthetically added, then reads are generated and mappability is assessed as before – by measuring the number of reads that map correctly. However, a read is now considered to map correctly if and only if the read maps to the correct location, and the CIGAR string indicates the correct type of variation.

This process is illustrated in Figure 5.3. Note that all mapped reads, even those with mapping quality zero, are considered as candidates for correct mapping.

The regions of greatest interest are those that have good mappability without any variation, but low mappability when a variation is present. These regions are sensitive to the presence of variations and reveal alignment biases that arise when inferring differences between a donor and reference genome. This approach eliminates regions that already have poor mappability due to the reference genome, and highlights regions that are affected by variation in the donor.

By comparing mappability results with and without the variation, the effect of the variation can be measured. A significant difference indicates a location where variant detection is difficult or impossible. By observing problematic loci across a number of genomes, generalised circumstances that cause bias can be determined.

A common post-alignment approach to indel calling is local reassembly (see Section 3.2.3). Since aligners consider each read independently, an optimal alignment for each read can result in contradictory alignments. We used *IndelRealigner* from the *GATK suite* (DePristo et al., 2011) to assess whether realignment could resolve poorly aligned indels.

5.2 Results

In this section, we present the results obtained when applying point mutations to the reference genome using the simulation framework. Next, we extend this analysis to include indels. Finally, we present the results obtained from applying mappability analysis to several genomes.

5.2.1 SNVs

SNVs are a widely studied and important component of the analysis of genome variations. Here, we study the capacity of aligners to accurately map reads, and to produce evidence that enables the discovery of SNVs, with respect to a number of influencing parameters. Simulations of read alignment and SNV prediction generated the following results:

- Any repeated sequence longer than the read length will either not be mapped to, or will be mapped to with zero mapping quality. This is a critical limitation of alignment, as DNA sequences typically consist of a non-trivial proportion of repeated sequences that are longer than current read lengths (see Section 4.2.1). Increased read lengths improve both alignment accuracy and SNV prediction accuracy, primarily by resolving repeats. With no sequencing errors, the number of unmapped reads is directly proportional to the number of repeated sequences that are equal to or longer than the read length.

- The likelihood of a read being mapped is proportional to the probability of finding a seed in the read, which decreases as differences between the read and the correct location accumulate. Sequencing errors have the same effect. Reduction of the seed length enables increasingly divergent regions to be mapped to the genome: with a read length of 100 bp, a seed length of 20 bp is effective with a mutation rate of up to 4%, while a seed length of 8 bp is effective at a mutation rate of up to 8%.

- Although higher coverage does not improve alignment accuracy, it enables variations to be predicted with greater confidence. In particular, if variations are covered by two or more reads, false positives due to uncorrelated errors and false negatives due to lack of coverage are largely eliminated. Ignoring coverage biases arising due to the sequencing process, 20x sequencing coverage is sufficient to predict homozygous SNVs by providing at least three reads of coverage across the genome.

Mutation Rate

Section 2.3 introduced the seeding strategy that is currently employed by aligners to find potential alignments without requiring an exhaustive search for approximate matches. For a read to match a location, there must exist a seed within the read that aligns exactly to a location without mismatches. Once a list of seeds is found, the aligner extends these seeds using an approximate matching algorithm

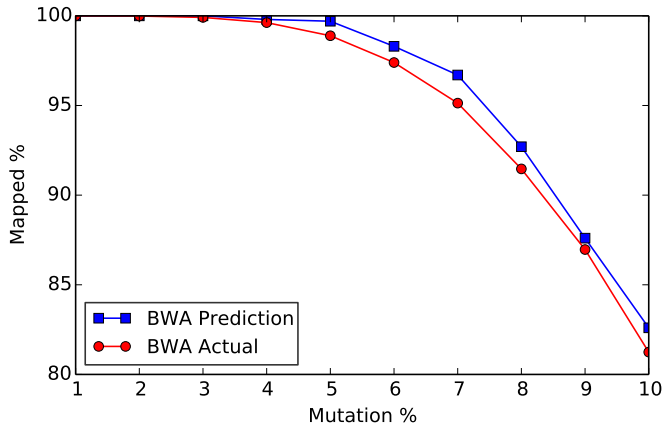

Figure 5.4: Mapping rate as a function of mutation rate. We compare the observed proportion of reads mapped to E. coli *to the predicted proportion of reads containing seeds of length 19 bp. The strong correlation between the two curves illustrates that most reads are not mapped because they do not contain a seed.*

to find the best approximate match for the read. If no seed matches the reference, the read will not be mapped.

Given a random distribution of mutations, the probability of a seed existing for any given read can be calculated by simulation (non-trivial approximate methods are described by Feller (1968)). Figure 5.4 illustrates the correlation between the expected mapping rate for a seed length of 19 bp, and the actual mapping rate achieved by *BWA-MEM*. The area above the prediction indicates reads that are not mapped due to not containing a seed, whereas the area between the curves indicates reads that are not mapped for some other reason. Most reads are not mapped because the read does not contain a seed.

BWA-MEM by default has a minimum seed length of 19 bp, while *Bowtie 2* by default has a minimum seed length of 20 bp. A direct consequence of this difference is that, at high mutation rates, *BWA-MEM* by default is more likely to map the read because it has a lower required length for an exactly matching region within the read. To illustrate this, we introduced mutations at fixed intervals across the genome, and observed that reads are only mapped when the interval between mutations is greater than the seed length, otherwise no exactly matching seed can be found. *Bowtie 2's* default settings include additional heuristics that result in higher than expected unmapped reads: with exhaustive search settings only the effect of the seed length remains (see Figure 5.5).

A solution to this problem is to reduce the seed length, but seeds are a necessary heuristic because it is infeasible to perform exhaustive edit-distance algorithms such as Smith-Waterman on large genomes with large numbers of reads (see Section 2.3). Figure 5.6 illustrates that reducing the seed length is primarily beneficial at high levels of variation from the reference: the seed length determines the maximum level of divergence that will be aligned.

A key point is that mutation rates vary considerably across a genome, and regions that exceed the

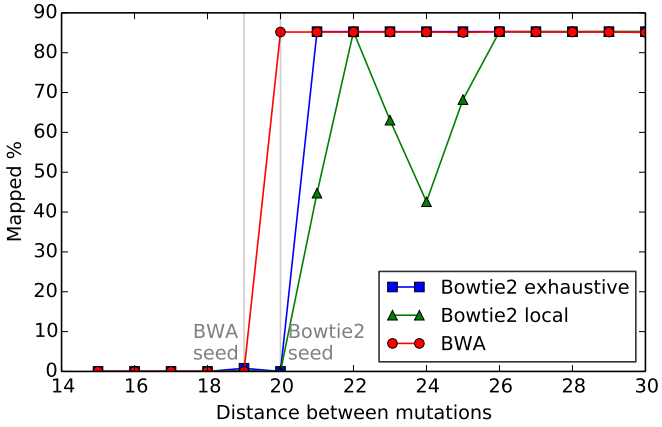

Figure 5.5: Mapping rate as a function of mutation period. By introducing mutations that are a fixed distance apart, the effect of the seed length can be observed. No reads are mapped if the distance between mutations is less than the seed length.

maximum level of divergence will not be aligned. Consequently, seed-based alignment inherently introduces a bias because all mapped regions must contain a seed, and sequence with sufficient variation will not contain a seed.

A primary goal of alignment is to discover variations and, in particular, SNVs. Variant callers assess the evidence presented by all aligned reads that overlap the variant of interest (see Section 2.3). Since this study does not consider the ability of variant callers to assess this evidence, a more useful measurement is the proportion of variations for which there is *no* evidence in the alignment. If no correctly aligned read covers a variation, then there is no evidence indicating that variation. Figure 5.7 shows the proportion of variations with at least one spanning read, at varying levels of mutation. Redundancy of coverage enables most variations to be discovered until a threshold is reached (approximately 8% mutation rate), after which, variations are lost approximately linearly.

Errors

Sequencing errors and variations have the same effect on individual reads. Errors introduce noise into reads and effectively increase the edit distance from the read to the correct location on the reference. Errors reduce the likelihood of the read aligning correctly by reducing the minimum seed length required to map the read. When considering the ability to map reads, addition of the error rate to the mutation rate provides the effective mutation rate as seen by the aligner.

In areas of low coverage, the existence of errors limits the ability to confidently predict variations. Assuming a variation rate v and error rate ϵ, in areas covered by a single read, the false positive rate is $\frac{\epsilon}{v+\epsilon}$, because errors cannot be distinguished from variations. Variation rates and error rates are typically of similar orders, suggesting that variation prediction from a single read is unreliable.

The probability of errors falsely suggesting a variation rapidly falls as coverage increases. Not

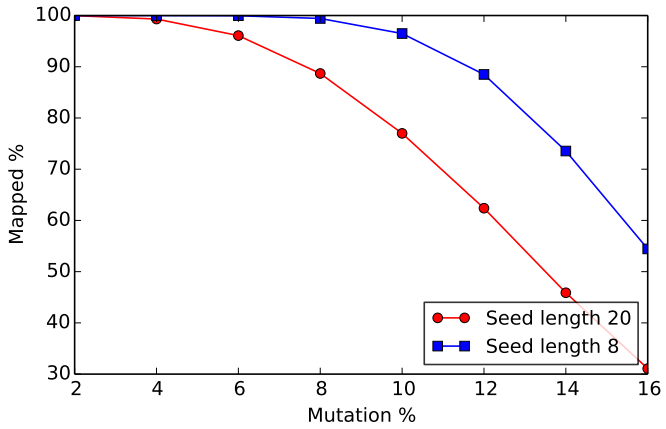

Figure 5.6: Mapping rate as a function of mutation rate with different seed lengths. We aligned reads containing randomly distributed mutations to E. coli *with* BWA-MEM *and observed the proportion of mapped reads. A shorter seed length is more effective with divergent sequences.*

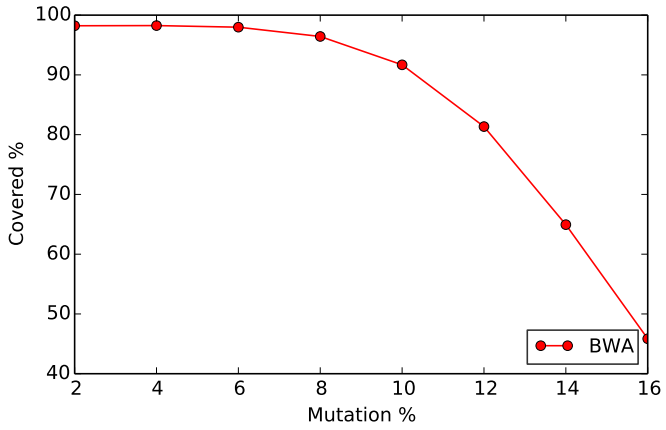

Figure 5.7: Proportion of variations with evidence indicating their existence at 20x Poisson distributed sequencing coverage. An increasing number of variations are not covered at high mutation rates.

Coverage	1% error	2% error	10% error
1	1 in 2	2 in 3	10 in 11
2	1 in 800	1 in 300	1 in 44
3	1 in 320,000	1 in 60,000	1 in 1760
10	1 in 10^{23}	1 in 10^{20}	1 in 10^{14}

Table 5.1: Expected incidence of false positives due to errors with a 1% variation rate between reference and donor. Errors and variations are assumed to be randomly distributed point mutations.

only must the position of the error match, but, to suggest consistent evidence for a variation, the base must also match. Assuming errors are randomly distributed, equally probable, and independent, and the variant is homozygous, each additional covering read has probability $\frac{\epsilon}{4}$ of matching the base of the first erroneous read. Assuming that a false positive occurs only if every read covering a position falsely and consistently indicates a variation, then the false positive rate for coverage c is:

$$FP = \frac{\epsilon}{v + \epsilon} \left(\frac{\epsilon}{4}\right)^{c-1}$$

Using this formula, Table 5.1 illustrates the expected false positive rate for homozygous variations as a function of error rate and coverage, with a variation rate of 1%. The probability of an error being misinterpreted as a homozygous variant rapidly drops as coverage increases. If the organism is diploid or polyploidy then variants can be heterozygous and more reads are required to confidently predict a variation against a background error rate. Although this simple model makes a number of assumptions, Bentley et al. (2008) similarly found that two spanning reads was sufficient for detecting most homozygous SNVs using real data.

Read Length

Reads that originate entirely from within a repeated sequence cannot be unambiguously mapped. However, as the read length increases, shorter repeated sequences no longer cause ambiguity. Once the read length is longer than the longest repeated sequence, all loci become unambiguous. Figure 5.8 illustrates the relationship between repeated sequence and the read length for *E. coli* K-12 MG1655. More reads become unambiguously mappable as the read length increases. Hence mapping rates and coverage increase up to the limit of the proportion of the genome that is uniquely mappable.

The expected improvement as read lengths increase is genome-specific and dependent on the distribution of the lengths of repeated subsequences. For example, *E. coli* K-12 MG1655's longest repeated subsequence is 2815 bp, implying that a read length of greater than 2815 bp is required to unambiguously resolve all reads. In contrast, *HIV-2's* longest repeat is 854 bp and *P. falciparum's* longest repeat is 11 909 bp. While the read length is less than the longest repeat, increasing read length improves alignment accuracy and variation coverage. Longer reads contain more seeds which increases the probability of finding a matching seed.

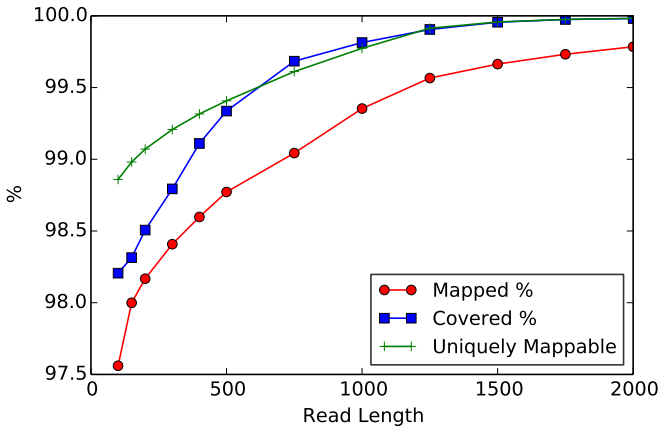

Figure 5.8: Effect of read length on mapped reads and covered variations relative to unique sequence. As read length increases, the proportion of mapped reads and the proportion of variations with at least one spanning read increases. Both of these values approach the proportion of the genome that can be unambiguously aligned to.

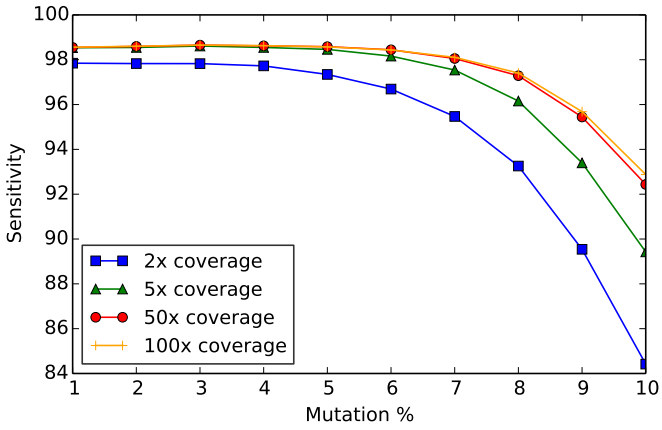

Figure 5.9: Relationship between mutation rate, sensitivity, and coverage. Increased coverage is more beneficial at high mutation rates.

Coverage

Coverage does not affect mapping accuracy since each read is aligned independently. However, coverage enables more confident predictions by providing more supporting evidence for a hypothesis. For example, confidence in the existence of a variation given a single supporting read will be low because there is a relatively high probability that differences are due to sequencing errors. Confidence rapidly increases as more reads support a variation (see Table 5.1).

Consequently, increased coverage helps compensate for errors: if errors are independent, then with more coverage the consensus is more likely to be correct. Figure 5.9 illustrates that although higher coverage is always beneficial, there is more benefit at higher variation rates. The benefits of

additional coverage rapidly diminish: once a homozygous variation is covered by at least three reads there is minimal benefit to additional coverage.

The sequencing coverage required to cover the entire genome under a Poisson distribution model was calculated by simulation, and is dependent on the genome length, read length, and number of reads. The average sequencing depth required to cover *E. coli* K-12 MG1655 with a read length of 100 bp is 14x for at least one read, 17x for at least two reads, and 20x for at least three reads. This approximately matches the experimental results of Bentley et al. (2008), who found that 15x sequencing coverage was sufficient for detecting most homozygous SNVs.

5.2.2 Indels

By applying the simulation and mappability framework to indels, we found the following:

- All the limitations discovered with SNVs also apply to indels: repeated sequences longer than the read length create ambiguity, and unmapped reads occur when all seeds in a read contain variation or error.

- Longer indels cannot be directly inferred from aligned reads. At a read length of 100 bp, there is no direct evidence for insertions longer than 30 bp or deletions longer than 45 bp. In general, for an insertion to be predicted directly from the alignment, the read length must be at least triple the length of the insertion; for a deletion the read length must be at least double. This relationship between read length and indel length is a direct reflection of the scoring system used to assess an alignment: the cost of a gap increases linearly with the length of the gap, until the score falls below some threshold that prevents the read from being aligned.

- Although shorter indels (less than 10bp) are generally detectable, this is dependant on the structure of the genome, the location of the indel in the genome, and the content of the variation relative to the surrounding content. STRs that extend beyond the start or end of a read can result in an alignment that does not suggest any variation, or may suggest an alternative incorrect variation. This is more likely if the indel length is a multiple of the repeat length and the indel matches part of the repeat, because this allows a perfect alignment without the variation. This scenario is particularly problematic when the variation is near the start or end of a read, because the STR need only extend to the edge of the read for the variation to be lost. STRs tend to be longer and more frequent in eukaryotes (Richard et al., 2008).

- Near-repeats are common in genomes. An indel reduces the amount of matchable sequence which increases the likelihood of multiple matching loci: although the read length remains the same, matching content from the reference is reduced by the indel. Deletions are particularly prone to this issue because a deletion may remove the subsequence that differentiates the two similar regions. As with exact repeats, increasing the read length helps resolve ambiguity between similar sequences.

	Indel Length			
Read Length	10	20	30	50
100	30	55	333	-
200	24	29	37	86
300	22	25	29	41
1000	20	21	22	23

Table 5.2: Required sequencing depth with a given read length to detect indels up to a given length.

- In addition to resolving repeated sequence, increasing the read length also reduces the effect of clipping and STRs. Although the number of bases affected by the variation remains the same, this represents a smaller proportion of the read.

- As with SNVs, confidence in a predicted variation requires a minimum number of reads spanning the variation. We observed that indels up to 10 bp resulted in up to 33% of reads being discarded on average, with higher rates in areas of near-repeats. Consequently, to detect most indels up to 10 bp, coverage must be increased by 50% relative to homozygous SNV detection. For *E. coli* K-12 MG1655, this corresponds to 26x to 30x sequencing depth to retain a minimum coverage of 2x to 3x respectively. This result can be generalised: the required sequencing depth D for a read length r and maximum indel length l to achieve a minimum coverage of 3x across the genome is:

  $$D = \frac{20r}{r - 4 - 3l} \quad \text{for } l > 2, r > 4 + 3l$$

  We derive this formula from our observation that the average proportion of reads affected by an insertion is $\frac{4+3l}{r}$ (see Section 5.2.3), and our calculation of 20x coverage required to recover SNVs (see Section 5.2.1). If a fraction f of reads is affected, then coverage $\frac{D}{1-f}$ is required to maintain coverage D. We combine these facts to calculate the required indel coverage. Example coverage requirements based on this formula are shown in Table 5.2.

Indel Length and Read Length

To evaluate the effect of the length of the indel, we synthetically added variations to *E. coli*, and artificially generated reads. To limit interference between variations, we enforced a minimum distance between variations equal to the read length. We aligned reads using *BWA-MEM*. As with SNVs, we focus on detecting whether *any* evidence for the variation is present, even when this does not appear to be the most likely outcome.

Table 5.3 shows the longest indel that has direct supporting evidence in the aligned reads. With default settings, *BWA-MEM* is strictly superior to *Bowtie 2* with respect to sensitivity when detecting long indels. The relationship between indel length and read length is approximately linear and in general, to detect an indel of length l, a read length of at least $3l$ is required. Although *Bowtie 2*

	Max Insertion Length		Max Deletion Length	
Read Length	*BWA-MEM*	*Bowtie 2*	*BWA-MEM*	*Bowtie 2*
100	30	20	45	30
200	60	45	95	50
300	95	70	145	70
400	130	95	180	95
500	160	120	230	110
750	245	185	255	165
1000	330	245	365	220

Table 5.3: Longest insertions and deletions that have supporting evidence in aligned reads. Simulated indels with lengths increasing in multiples of five were added to 100 bp reads, then aligned to E. coli.

	Insertion		Deletion	
Indel Length	Mean Loss	Maximum Loss	Mean Loss	Maximum Loss
1	6.7	15	5.8	47
2	9.6	18	10.5	56
3	12.6	21	12.6	83
4	15.5	25	14.5	86
5	18.5	27	16.4	86
6	21.5	53	18.5	86
7	24.4	34	20.4	86
8	27.4	54	22.4	86
9	30.4	39	24.4	86
10	33.3	42	26.3	87

Table 5.4: Average and maximum percentage of reads affected at a location when an indel is synthetically added to E. coli *and aligned with* BWA-MEM. *We measure the number of reads affected by comparing the number of spanning reads that align correctly with and without the variation present.*

appears to be less sensitive than *BWA-MEM*, this is due to different default scoring systems. *BWA-MEM's* and *Bowtie 2's* gap extension penalties are 1 and 3 respectively (see Table 2.1), which prevents *Bowtie 2* from aligning long indels by default. Note that by default *BWA-MEM* does not detect indels longer than 100 bp; this was overridden with the `-w` parameter in our experiments.

At all indel lengths, *BWA-MEM* more accurately detects deletions compared to insertions, which gives rise to an asymmetry: the same indel may be detectable when it is a deletion, but not detectable when it is an insertion. Schröder et al. (2015) exploit this difference by augmenting the reference with all known insertions, thus maximising the incidence of deletions relative to insertions.

5.2.3 Indel Mappability

We have determined the upper limit of the length of detectable indels. We now investigate in further detail systematic biases influencing the detection of shorter indels.

We applied mappability analyses to the reference genomes of *HIV-2*, *E. coli* K-12 MG1655, *D. melanogaster* chromosome 2L, and *H. sapiens* chromosome 21. Indels of length 1 bp to 10 bp were analysed, as well as SNVs. By investigating loci with significant loss of mappability, we identified three commonly occurring biases that limit indel detection: clipping, STRs, and similar sequence.

Clipping

When a variation occurs close to either end of a read, the aligner often clips the variation off the read and aligns the rest of the read. This is a consequence of the Smith-Waterman algorithm that does not require the entire read to be aligned to the reference. Clipping reduces the number of reads that directly suggest the indel. This implies that for the indel to be detected, other reads must span the variation without clipping. In the majority of loci, clipping is the only factor that reduces mappability. Clipping rates were consistent across all genomes tested.

Provided that the read length is at least triple the length of the variation, and coverage is high enough to tolerate the loss of some reads through clipping, indels can be confidently predicted. For example, 10 bp insertions on average lose 33% of spanning reads due to clipping, which suggests that 50% more reads are required on average to maintain the same coverage.

Increasing read length reduces the effect of clipping. The same number of bases are clipped regardless of the read length. Consequently, doubling the read length halves the effect of clipping. Since clipping is the main obstacle to detecting most indels, this linear relationship is reflected in the maximum detectable indel shown in Table 5.3.

STRs

STRs are present at the majority of loci that exhibit significant loss of mappability. STRs are common in eukaryotes: approximately 7% of the human genome is covered by 260 000 STRs (Richard et al., 2008). Our results reflect this: when analysing *HIV-2* and *E. coli*, STRs posed limited problems in indels up to 10 bp (see Table 5.4), but STRs can eliminate all evidence for insertions as short as 2 bp in *H. sapiens*, and deletions as short as 4 bp in *P. falciparum*.

STRs cause problems when detecting both insertions and deletions and are manifested in a wide variety of mechanisms. Indels are more likely to be lost if the STR length is a multiple of the indel length, or if the content of the STR matches that of the indel. However, neither of these conditions are necessary. The STR also need not be an exactly repeating sequence. If the cost of including the true variation in the alignment exceeds that of aligning the variation to the STR, the alignment will fail to suggest the variation.

An example of an incorrect alignment is illustrated in Figure 5.10, which occurs on *H. sapiens* chromosome 21 at position 34 950 000, visualised using *IGV* (Robinson et al., 2011). The reference sequence consists of a series of repeated “AC” bases and the donor contains an inserted “A”, but this variant is incorrectly interpreted as a deleted “C”. By deleting “C” the same sequence as the reference

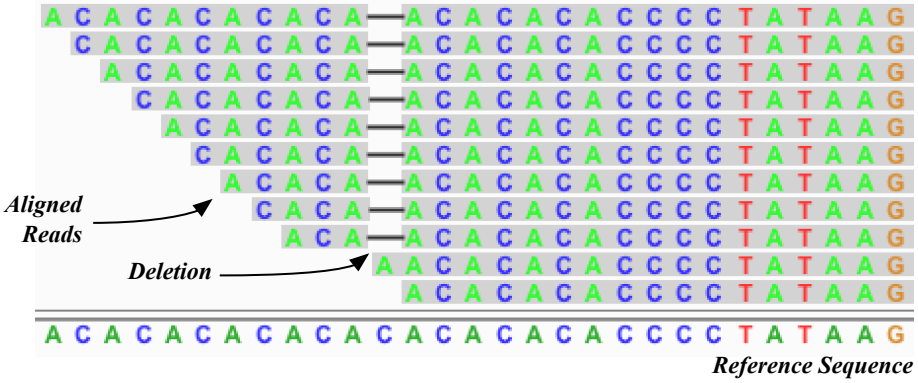

Figure 5.10: This STR causes an insertion to be interpreted as a deletion. By effectively deleting a repetitive unit the sequence remains identical to the reference over the length of the read.

is obtained, with one less repeated "AC", which in this example can be moved beyond the start of the read.

In general, if an insertion contains part of an STR, the rest of the STR can be deleted and, if the repetition extends beyond the end of the read, the missing STR can be moved beyond the end of the read.

Similarly, Figure 5.11 illustrates a 2 bp insertion on *P. falciparum* chromosome 2 at position 642 300. Many of the reads map with high confidence yet suggest the wrong variation. Figure 5.12 shows that variant inference can also be sensitive to the starting position of the read. A 6 bp STR causes *BWA-MEM* to align the insertion to part of the repeat and delete the remainder. Consequently, the alignment suggests that one (or more) STRs are deleted and fails to indicate the insertion. *IndelRealigner* does not resolve the discordant reads, instead suggesting an incorrect variation.

If an insertion approximately matches an STR then the aligner may overlay the insertion on to the STR (see Figure 5.13). Instead of an insertion, the predicted variation is the difference between the inserted sequence and the STR.

Each of these examples of failure follow a common pattern: the presence of repeated sequence enables sequence to be deleted without impacting on the alignment, and the cost of aligning to the repeat is at least as low as the cost of including the novel content. If both of these properties hold then an incorrect alignment is likely to arise.

Deletions are also prone to bias caused by STRs. Figure 5.14 shows the alignment across a 4 bp deletion on *P. falciparum* chromosome 2 at position 435 600. Despite the repeated region being significantly shorter than the read length, the alignment presents no evidence for the deletion, instead suggesting a point mutation 16 bp away from the true variation. The repeating unit "ATAT" is followed by repetitions of "ATGT". Effectively, the last "ATAT" overlaps with the first "ATGT", which enables the deletion to be represented in the aligned reads with a single point mutation. A single point mutation is penalised less than a deletion so is the preferred, albeit incorrect, solution.

Local realignment is commonly included in many bioinformatics pipelines. We found that local

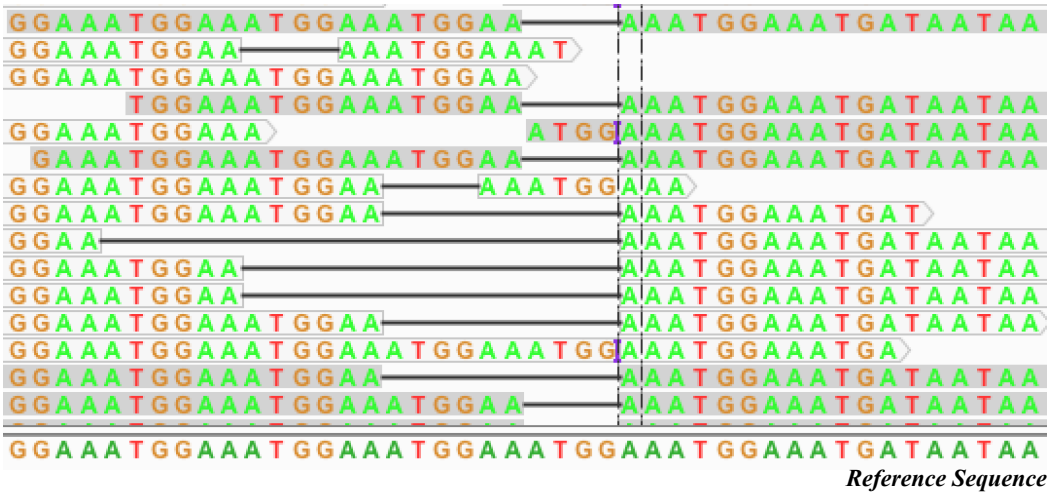

Figure 5.11: A 2 bp insertion causes a variety of incorrect interpretations, including on reads aligned with high confidence (shaded grey), due to the existence of multiple alternative explanations. A 6 bp STR is responsible.

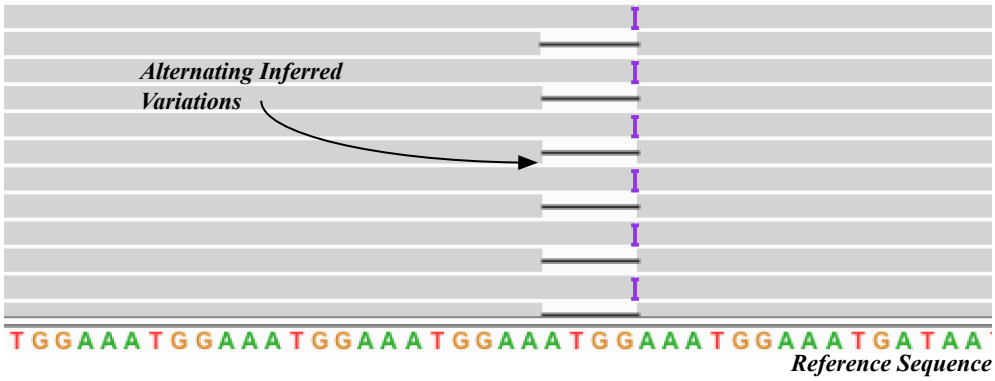

Figure 5.12: These alignments are sorted by start position. Small differences in the read location cause differing interpretations when STRs are present.

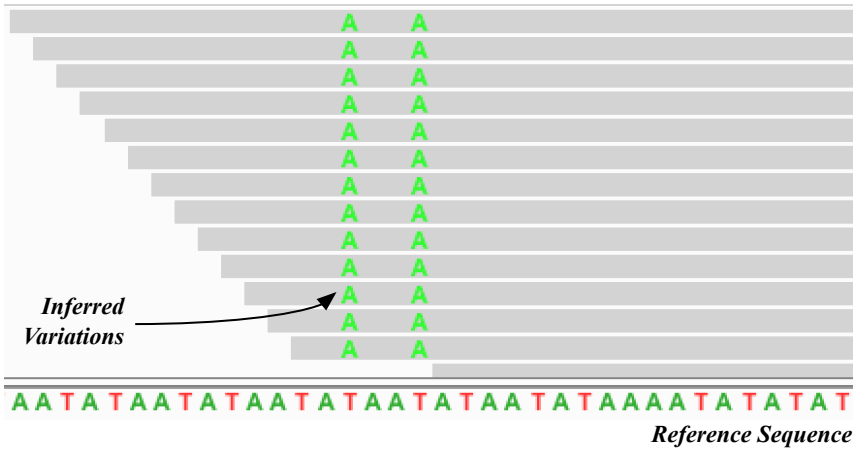

Figure 5.13: A 5 bp insertion is instead matched to the 5 bp STR with two mismatches. The insertion is lost and the two SNVs are called with high confidence.

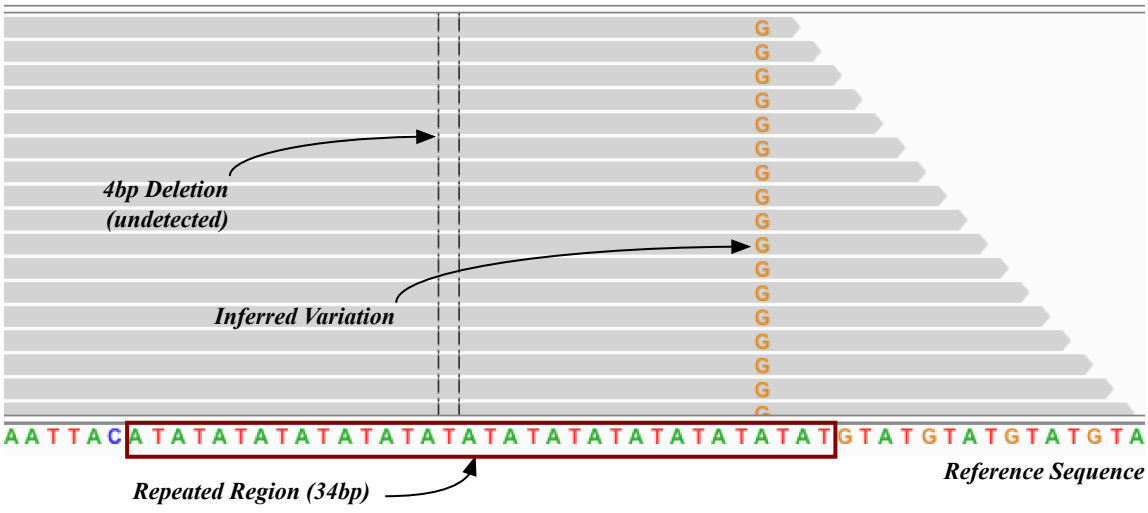

Figure 5.14: A 4 bp deletion matches the period of the STR. Rather than suggesting the deletion, the alignment suggests a point mutation 16 bp away from the true variation.

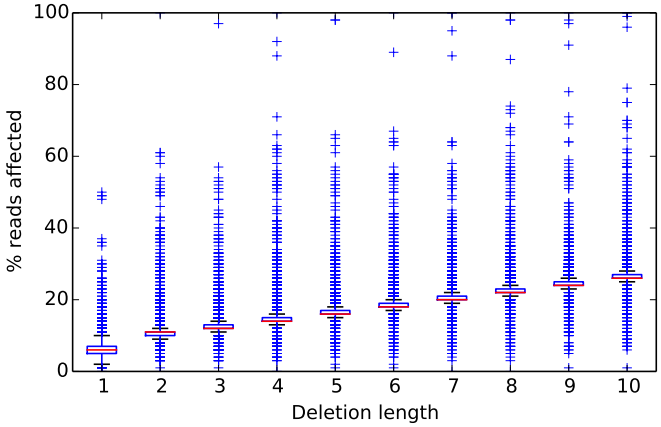

Figure 5.15: Boxplot illustrating the distribution of reads affected by deletions across 20 000 H. sapiens *loci.*

realignment was partially successful when resolving ambiguity created by STRs. Short STRs, such as those found in *HIV-2* and *E. coli*, were usually resolved by local realignment. In contrast, longer STRs, such as those found in *P. falciparum*, *H. sapiens*, and *D. melanogaster*, were typically not resolved by this step. Local realignment also sometimes exacerbates results by unifying the aligned reads to present a consensus, all indicating the same incorrect variation.

Similar Sequence

Genomes, particularly eukaryotes, consist of significant amounts of repeated and near-repeated sequence due to paralogs, pseudogenes, and transposable elements. Even though a region may be unique across the genome and hence highly mappable, many regions are very similar (see Section 4.2.1). Consequently, a unique region may only require a small perturbation to be indistinguishable from another region, thus creating ambiguity.

A single base substitution or deletion at 24 031 000 on *H. sapiens* chromosome 21 results in the loss of 42% of reads overlapping this position, and most of the remaining reads are given mapping quality zero. This change creates ambiguity because there are multiple regions on this genome that differ only by this base. If this difference is changed or deleted, the regions become indistinguishable.

Prevalence

The prevalence of bias for a given reference is dependent on the nature of the reference and the differences between the reference and the donor. *HIV-2* and *E. coli* are less affected by STRs than *H. sapiens*, *P. falciparum*, or *D. melanogaster*, but significant differences also exist between the eukaryotes tested. Richard et al. (2008) suggest that STR prevalence varies significantly between strains of the same species.

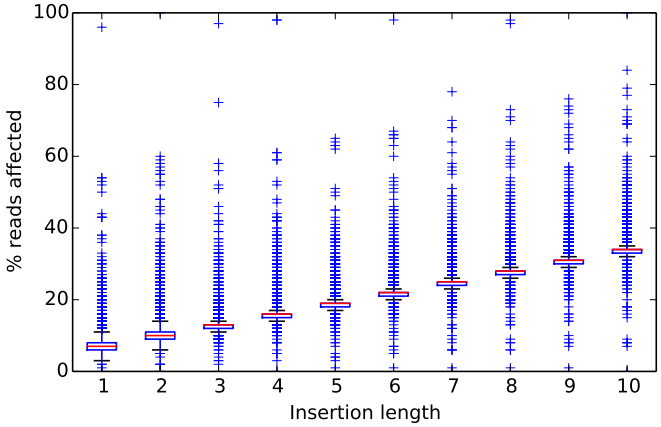

Figure 5.16: Boxplot illustrating the distribution of reads affected by insertions across 20 000 H. sapiens *loci.*

To measure the distribution of bias caused by short variations, we uniformly sampled 20 000 loci across *H. sapiens* chromosome 21 and calculated the mappability at each location relative to the variation. Figure 5.15 and Figure 5.16 illustrate the distribution of the proportion of reads affected by different length indels. The majority of loci are affected similarly: 95% of samples fall into a range of 5%. However, a small number of loci are critically affected. Consequently, although increased coverage helps to offset the loss of reads, at some loci, a variation cannot be recovered regardless of the level of coverage.

Based on these results, the mean proportion of reads affected by a variation is linear with respect to the variation length. For a read length of 100 bp, if i is the insertion length and d is the deletion length, then the percentage p of reads affected is on average:

$$\begin{aligned} p &= 4 + 3 \times i \\ p &= 7 + 2 \times d \end{aligned}$$

This linear relationship predicts 100% read loss at an insertion length of 32, and a deletion length of 46, which closely corresponds to the limits to variation prediction observed in Table 5.3.

5.3 Summary

We have illustrated limitations and biases that arise during alignment when inferring the presence of short variations, with a focus on biases that occur only when the variation is present.

STRs lead to incorrect variation inference, even when realignment techniques are employed. The effect is dependent on the length of the indel and its content. Correct inference of STRs is critical in some applications (see Section 2.1); these results suggest that alternative approaches to STR detection,

such as *lobSTR* (Gymrek et al., 2012), are required.

Repeated and nearly repeated sequence that is longer than the read length causes ambiguity and sensitivity to small changes. Heuristics employed by aligners assume a minimum region of exactly matching sequence; at high levels of variation this assumption breaks down. Many of the regions illustrated do not appear to be problematic with current evaluation measures such as mapping quality or mappability.

Increased read length helps resolve ambiguity, but analysis of repeat lengths in existing genomes suggests that repeats will be an issue for the foreseeable future. Increasing sequencing coverage also helps maintain sufficient depth even when many reads are lost due to biases, but this has a diminishing return. There are limits to the improvement attainable purely through increased coverage.

Chapter 6

Inferring Bias with Whole-Genome Mapping

We have identified circumstances in which biases arise when inferring the presence of short variations in a donor genome. Unless the donor and reference genomes are very similar, large-scale differences that are not spanned by any read are also likely. This includes rearrangements, insertions, and deletions at a genome-wide scale. Whole-genome alignment enables us to directly observe bias arising from the choice of reference and its associated differences. Specifically, we assess the capacity of a short read aligner to accurately reflect the donor genome when aligned to different reference genomes.

6.1 Method

Mauve (Darling et al., 2010) is a whole-genome alignment tool that generates a map of differences between two genomes in terms of rearrangements, gains, and losses. In a process similar to short read alignment, the *Mauve* algorithm first finds seeds of approximately matching subsequence on each genome, then extends these seeds to generate a set of anchors for each genome. These sets of anchors are then aligned across the genomes to enable the set of most likely rearrangements to be determined. A significant difference, relative to short read alignment, is that whole-genome alignment has access to both genomes, and is not limited by the read length. Consequently, seeds can be extended without limit, enabling accurate modeling of large-scale rearrangements. This mapping can be used to evaluate the effect of the choice of reference. The process is outlined in Figure 6.1 and consists of the following steps:

- An existing set of reads is independently aligned to the two genomes of interest.
- *Mauve* generates a high level mapping between the two genomes.
- Reads aligned to the reference genome are remapped to the donor via the Mauve mapping.

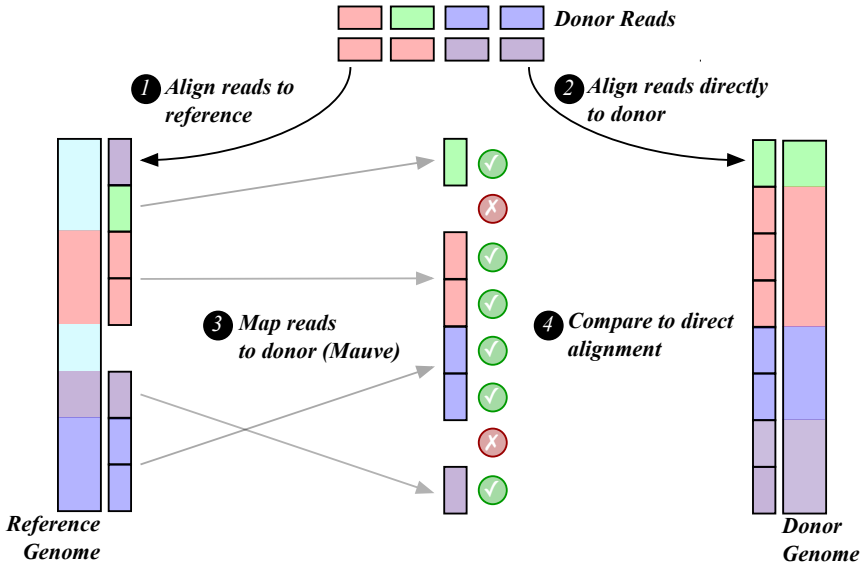

Figure 6.1: Whole-genome mapping is used to map aligned reads from one genome to another. Differences in the genomes can be associated with changes to the alignments.

- The impact of aligning to the reference genome is assessed by comparing each remapped read to the corresponding directly aligned read. This enables measurement of biases introduced by differences between the genomes.

Although the target of the direct alignment is referred to as the donor genome, the true source of the reads is not necessarily this genome. However, this experiment is most useful if the true donor is similar to the target genome. If this is the case then errors introduced by the mapping process are hypothetically representative of the bias introduced by using a reference genome that differs from the donor genome.

This hypothesis is subject to several assumptions. In particular, *Mauve's* mapping mechanism is a critical component – each rearrangement is assumed to be correct. We assume that novel content on the donor will not be the target of any mapping, and that novel content on the reference will not be the source of any mapping. If this is not the case, then errors introduced by the remapping process could be incorrectly categorised as bias.

In the case of real data, the true donor genome associated with the reads is not known. As a result, the direct alignment can be expected to contain biases, manifested as unmapped reads and incorrectly mapped reads (see Section 4.2.1). These are due to interactions between the mapping process and the structure of the donor genome, and not due to differences between the donor and reference genomes. Although aligners report reads that cannot be mapped, incorrectly mapped reads are difficult to identify when using real datasets and subsequently, cannot be corrected for. To minimise this bias, we selected *E. coli* K-12 MG1655 as the donor genome, an extremely well-studied strain of *E. coli* with a large number of available read sets. We selected a read set from the NCBI Short Read Archive

(Leinonen et al., 2010) with a high proportion of reads mapped with high quality, and with high coverage observed across the donor genome. The selected set of reads represents the donor genome as closely as possible.

Read errors are another confounding factor that can influence the apparent bias due to differences in the genomes. Errors increase the edit distance between the read and the correct alignment and are most likely to affect the outcome of a read when the read already contains significant variation. Since the reference genome is expected to vary more from the reads than the donor genome, errors are more likely to affect the alignment to the reference genome. We pre-processed reads with error correction software to minimise this effect.

Mapping of each read has the following possible outcomes:

- *The read successfully aligns directly to the donor, but fails to align to the reference.* This behaviour is expected when sequence exists on the donor genome that does not exist on the reference. Novel content on the donor is a fundamental limitation of alignment.

- *The read successfully aligns to the donor and the reference, but there is no mapping from the reference to the donor.* Either the mapping mechanism has failed or the read has been incorrectly placed on the reference genome. The advantage of whole-genome alignment is its ability to consider all surrounding sequence when assessing the suitability of an alignment. In contrast, aligning a short read is difficult when there are many similar candidate locations. A variation or error in the reference can result in alignment to an incorrect location. To assess whether this error is due to mapping or alignment, this type of error is further classified based on whether there is *any* mapping to the correct location.

  A further complication of the mapping process is that ambiguity arises when a read does not map to a contiguous region, for example, a read may span the edge of a rearrangement (a breakpoint).

- *The read aligns to the donor and the reference, but is mapped to an incorrect location.* Similarly, either the mapping mechanism has failed or the read has been incorrectly placed on one of the genomes. It is difficult to distinguish between these possible causes. However, the mapping process has access to the entire genome sequence, whereas the alignment algorithm only has access to a short subsequence. As a result, when results conflict, it is assumed that whole-genome alignment provides the correct location, and the reference genome alignment is incorrect.

- *The read does not align to the donor and aligns to the reference.* This indicates commonality between the read set and reference genome that does not exist in the donor genome. Reads falling into this category violate the assumption that the true donor genome closely matches the proxy donor genome. We expect very few reads to fall into this category.

- Finally, the alignment is considered to be correct if the location of the read as mapped by *Mauve* matches that of the directly aligned read, or both reads are unmapped.

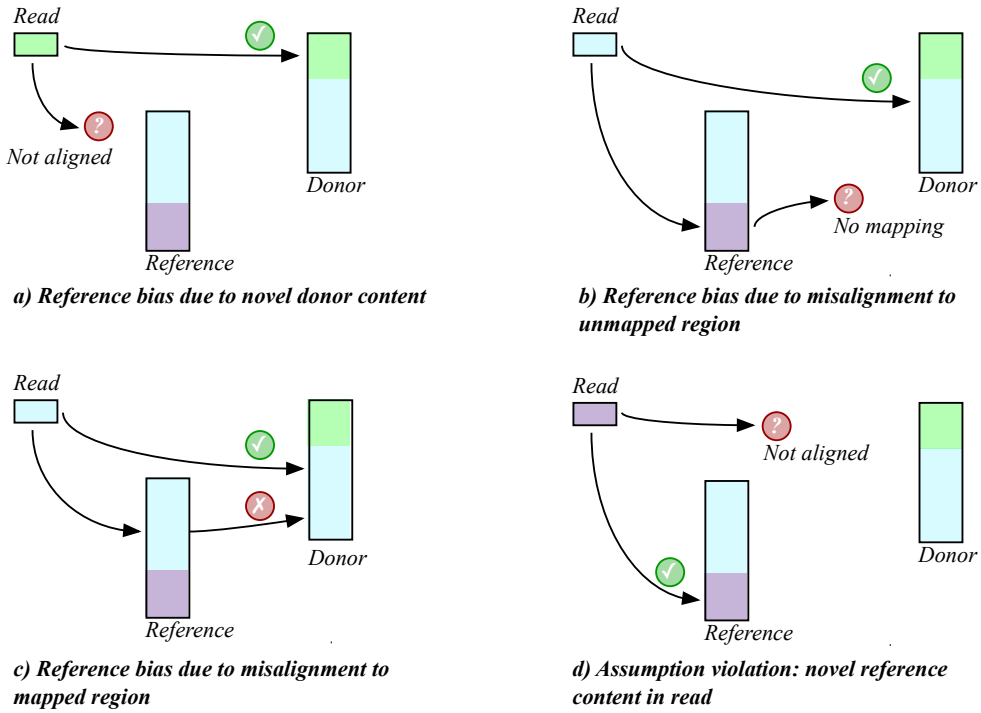

Figure 6.2: If the mapped read does not match the directly aligned read, the cause is either reference bias or the violation of an assumption of this experiment. Each possible outcome is illustrated here.

Figure 6.2 enumerates scenarios where the mapping process fails.

A key consideration when evaluating the absolute capacity for the donor genome to be accurately reconstructed is the coverage of the remapped alignment relative to the direct alignment. In particular, a significant proportion of the donor genome may be unreachable by *Mauve's* mapping process, either due to limitations of the mapping process, or the presence of novel content on the donor genome. This is an inevitable outcome if the reference genome is shorter than the donor genome, and provides an upper limit on the maximum proportion of the donor genome that can be covered after remapping.

Assuming that *Mauve's* mapping process is accurate, measuring the coverage of the correctly mapped reads against the directly aligned reads gives an indication of the bias caused by the choice of reference genome.

We used *BFC* (Li, 2015) for error correction, then aligned reads to both genomes with *BWA-MEM*. Given two genomes, *Mauve* generates an XMFA file which describes a mapping from each base in the reference genome to a location on the donor genome. Realignment from reference to donor considers each aligned read individually. If all bases on the read have a corresponding mapping in the XMFA file, then the read is aligned to the location and orientation provided by the XMFA mapping. If any base on the read does not have a mapping then the read is not aligned.

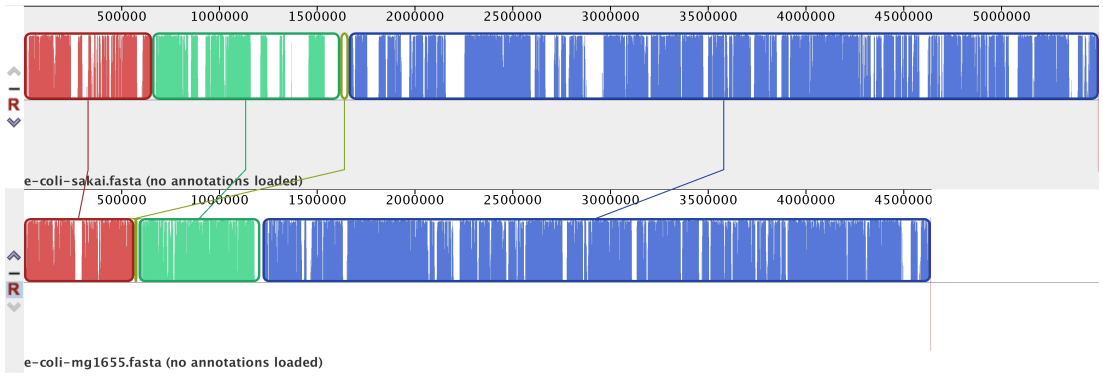

Figure 6.3: Mauve's *whole-genome mapping between* E. coli *O157:H7 Sakai (top) and* E. coli *K-12 MG1655 (bottom). There are four rearrangements, signified with coloured blocks. Inside each block is a similarity profile that indicates sequence conservation in that region.*

6.2 Results

E. coli is a model organism with a substantial number of sequenced strains. It is also considered to be highly mappable with low levels of repeats and STRs, an assumption confirmed experimentally (see Section 5.2.3). By minimising the effect of the reference, this enables us to highlight the effect of the *choice* of reference.

6.2.1 Examining Sources of Bias with Whole-Genome Mapping

To examine the impact of the choice of reference, we aligned reads originating from *E. coli* K-12 MG1655 (short read archive SRR892241) to both *E. coli* K-12 MG1655 (the donor) and *E. coli* O157:H7 str. Sakai (the reference). We then mapped the alignments from the reference to the donor using the mapping provided by *Mauve*.

First considering the direct alignment to the donor, 96.7% of the 7.1 million single-ended reads mapped successfully, covering 99.96% (4 639 621 bp) of the donor. Three contiguous regions have no coverage which total 2031 bp, the longest of which is 1177 bp. This demonstrates the high level of similarity between the donor genome and that represented by the read set.

In contrast, when aligning to the reference genome, 86.0% of reads mapped successfully, covering 75.8% (4 169 881 bp) of the reference. A total of 1 328 569 bp are not covered and the longest contiguous region of zero coverage is 45 023 bp. Since only 4.2 Mbp are covered on the reference genome, and the total length of the donor genome is 4.6 Mbp, at least 471 771 bp (10.2%) from the donor cannot be covered by the reads mapped to the reference, regardless of *Mauve's* mapping.

The whole-genome mapping calculated by *Mauve* (see Figure 6.3) found mappings between the two genomes for 4 135 095 bp. Similarly, this provides a lower bound of 506 557 bp (10.9%) for the minimum number of bases on the donor genome that will not be covered by reads mapped to the reference genome via whole-genome mapping.

Unsurprisingly, there is a strong correlation between the covered regions on the reference genome

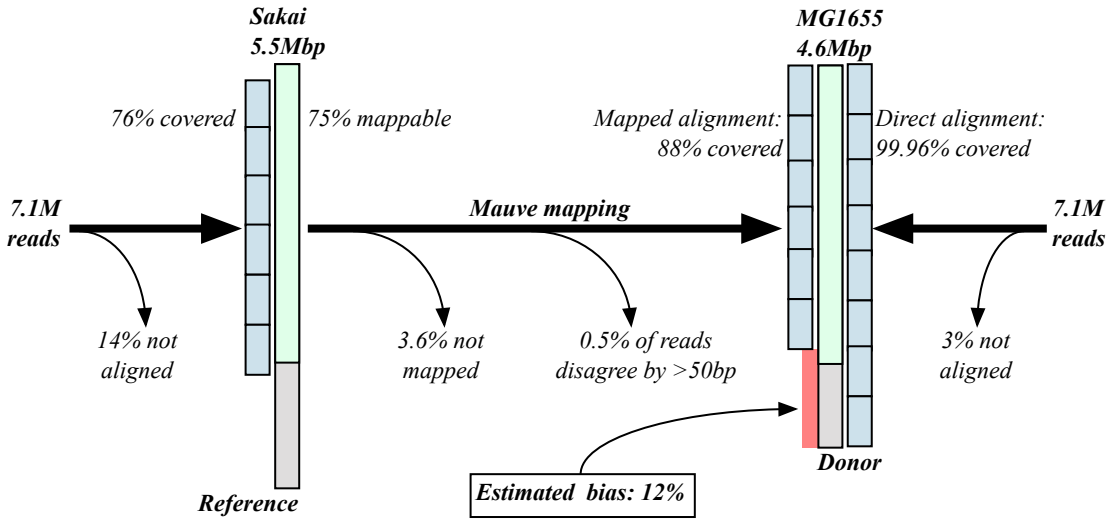

Figure 6.4: Measuring reference bias between two strains of E. coli. *We determined experimentally that in this instance, approximately 12% of the donor genome cannot be recovered, a loss that is entirely due to the choice of reference.*

and the regions that can be mapped to the donor genome. Of the reads that align successfully to the reference, 96.4% are completely contained in a mappable region, while 98.2% of reads have some overlap with a mappable region. After mapping reads from the reference genome to the donor genome, the resultant alignment failed to cover 556 671 bp (12.0%). The inferred genome contains 643 gaps and the longest contiguous region with zero coverage is 39 055 bp. We compared the aligned position of the remapped reads to the reads that were aligned directly to the donor, and found that 99.5% of remapped reads matched the correct location. Figure 6.4 summarises these results.

This analysis indicates that if a donor originating from *E. coli* K-12 MG1655 is aligned with *E. coli* O157:H7 Sakai as the reference genome, then bias arises from the following sources:

- *Sequence appearing in the donor that is not present in the reference:* 10.9% of the donor genome has no mapping to it from the reference genome. These regions contain sequence that is not present in the reference. Consequently, reads originating from these regions are not mapped to the reference. Without additional processing, these regions appear as deletions and any variations in these areas are undetectable.

- *Alignment to unrelated sequence:* 1.8% of reads align to regions on the reference with no mapping to the donor genome. This implies that these reads have incorrectly aligned to reference sequence that is not present in the donor sequence, or *Mauve* has failed to find the correct remapping. These off-target reads cover 76 922 bp (1.7%) of the reference sequence. Manual inspection reveals that approximately 20% of these reads are incorrect due to repeated sequence, 20% due to areas of high variability, and 60% due to *Mauve* considering the region of similarity too small to remap.

- *Incorrect alignment:* 0.5% of reads that are successfully remapped to the donor are placed at a location more than 50 bp from the correct location. These errors are highly clustered: 28% of mapping errors are clustered in less than 1% of the donor genome. Consequently, these errors can have a significant impact on a small number of specific areas. Manual inspection reveals that these areas contain repeated sequence in the donor genome that are unique in the reference genome, thus creating ambiguity. Erroneous reads cover 38 194 bp (0.8%) and coverage is up to 799. Since the average coverage of the entire realignment is 144, clustered incorrect alignments are likely to cause bias at specific locations by introducing reads at a coverage that is comparable to or greater than that provided by correct reads.

In the instance of aligning reads originating from *E. coli* K-12 MG1655 to *E. coli* O157:H7 Sakai, we find that 12.0% of the donor genome is ultimately unreachable. These unreachable areas cannot be reconstructed, and any variations in these regions cannot be recovered without additional analysis. The maximum additional detrimental impact of incorrect alignments is 0.8%, while perfect realignment that correctly aligns all off-target reads could theoretically improve this result by up to 1.7%. Consequently, we estimate that the bias introduced by the reference in this experiment is in the range of 10.3% to 12.8%. This is effectively the proportion of the donor genome that is not represented by any correctly aligned reads.

To assess the impact of paired reads, we repeated the experiment with the full set of paired reads. Bias improved by 0.2% to 11.8%, suggesting that although paired reads make a measurable difference, most of the bias remains.

6.2.2 Bias Across a Phylogeny

To explore the distribution of bias across a species, we calculated and assessed reference bias arising in 62 completed *E. coli* genomes by aligning reads originating from *E. coli* K-12 MG1655 to each alternative reference and then measuring bias. We found significant variability: bias ranges from 0.2% to 29.9% with a mean bias of 14.0%. Figure 6.5 illustrates the bias observed across this set of genomes, as well as the uncertainty associated with the measurement, and the pathogenicity assigned to each strain. Strains of similar pathogenicity tend to have similar levels of bias, and strains that exhibit high levels of bias tend to have greater measurement uncertainty.

Since pathogenicity is not necessarily always a good indicator of genome similarity, we generated a phylogenetic tree of this set of reference genomes. We generated genome distances with *andi* (Haubold et al., 2014) and clustered using *R's hclust* command. Figure 6.6 illustrates the relationship between reference bias and evolutionary distance. Although strains that are closely related to the donor genome exhibit low and predictable levels of bias (less than 5%), reference bias increases outside of this small cluster of strains and becomes increasingly difficult to predict. Related strains tend to exhibit similar levels of bias, but the level of bias that a group of strains exhibits is not obvious.

Two alignment measures that are strongly correlated with reference bias are the proportion of reads that are not mapped and the proportion of bases that are not covered. Unmapped reads are reported by

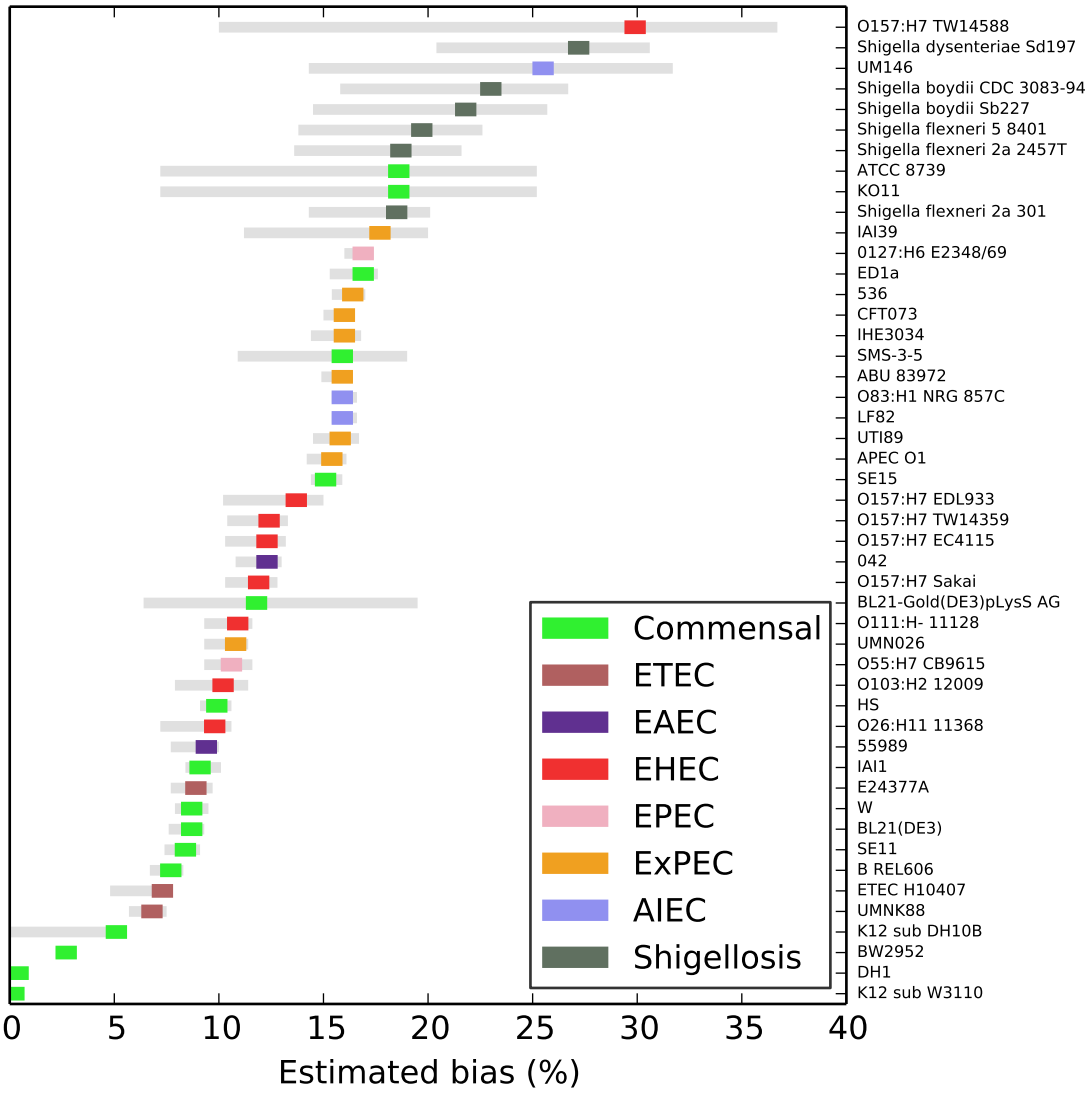

Figure 6.5: Calculated bias across 62 E. coli *strains for a read set originating from* E. coli *K-12 MG1655. Bars indicate measurement uncertainty while colours show the pathogenicity of the strain.*

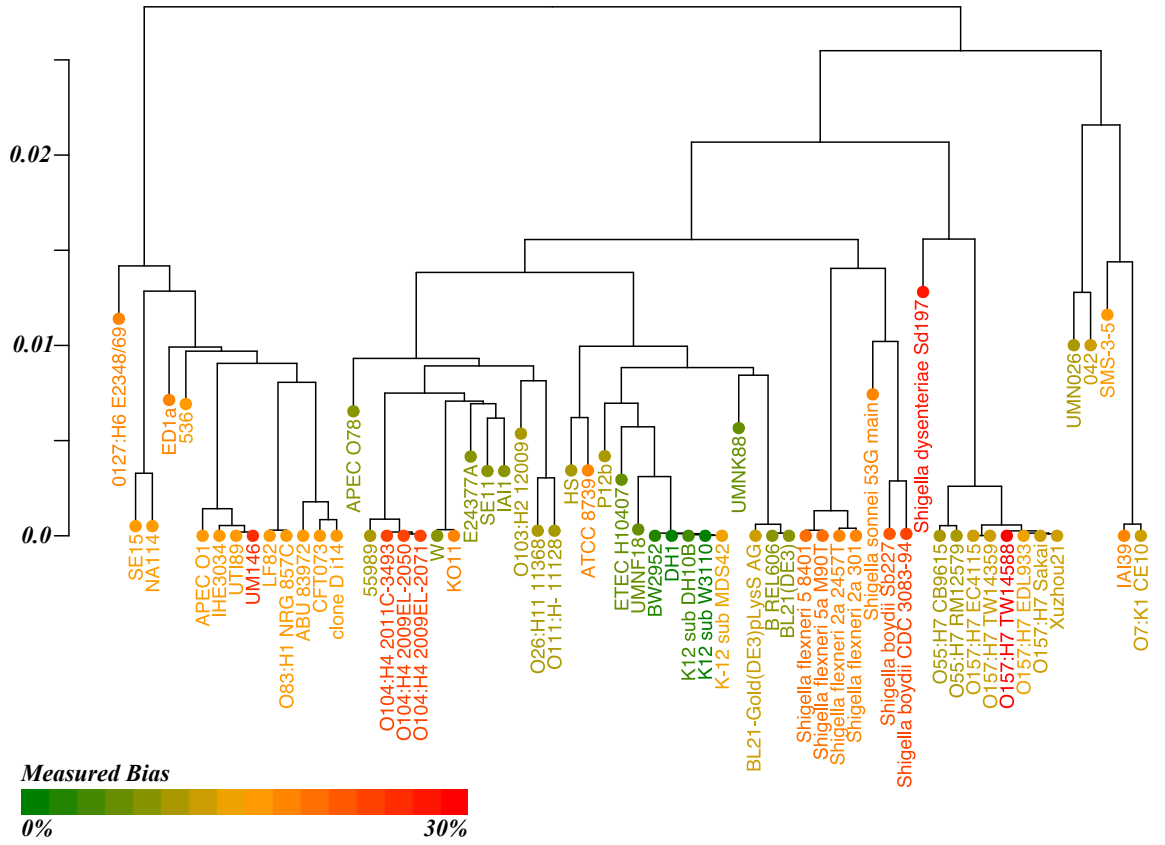

Figure 6.6: Calculated bias across 62 E. coli *strains for a read set originating from* E. coli *K-12 MG1655. A distance matrix was calculated using* andi *and clustered with* R. *Colour indicates the proportion of the donor genome that is lost when that* E. coli *strain is the reference genome.*

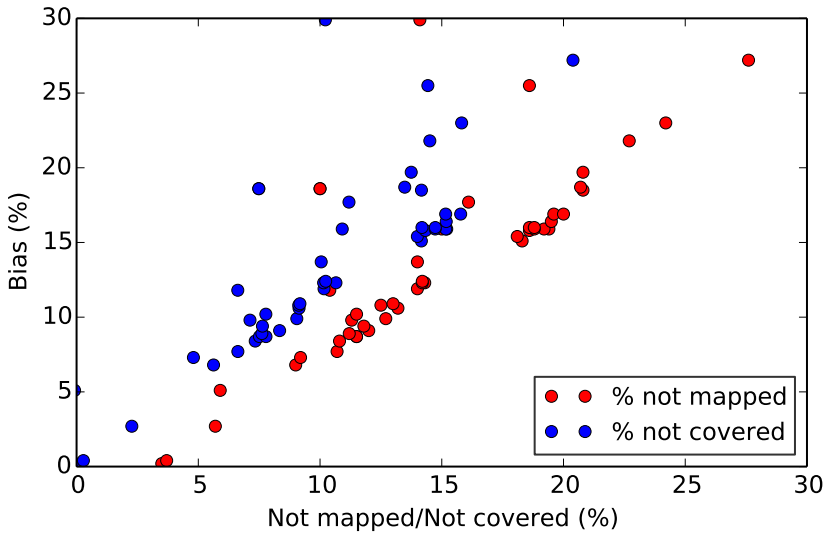

Figure 6.7: Correlation between calculated reference bias and unmapped reads and uncovered bases. Both measurements provide a lower bound for reference bias.

samtools and coverage was calculated using *bedtools*. Figure 6.7 illustrates the relationship between these two predictors and the observed reference bias. Low levels of unmapped reads and high levels of coverage are *necessary* to achieve low bias, but are not *sufficient*. Consequently, unmapped reads or coverage statistics enable a lower bound, but not an upper bound, of bias to be calculated.

6.3 Summary

In most cases, the majority of reference bias arises due to novel sequence that is present in the donor but not present in the reference. Consequently, reads originating from these regions are either not mapped, or mapped to alternative incorrect regions. The incidence of missing sequence necessarily increases as differences between reference and donor increase.

Bias varies from 0.2% to 29.9% across the 62 *E. coli* strains, suggesting that the choice of reference can have a significant impact on the result of an experiment. Most species have a very limited choice of reference genomes, which restricts the ability to reduce reference bias by selecting a different reference. *E. coli* is a model organism with a large number of reference genomes, theoretically enabling reference bias to be minimised by selecting the best reference. However, this choice is non-trivial, because estimating reference bias without prior knowledge of the donor genome is difficult. A lower bound can be calculated by considering coverage statistics and levels of unmapped reads on the reference genome, but other factors such as the complexity of the rearrangements between the two genomes can significantly influence bias. As a result, an upper bound is difficult to calculate without knowledge of the donor genome.

Chapter 7

Conclusion

We have identified and quantified bias arising from three main sources: the structure of the reference genome, small variations between the donor genome and the reference genome, and large-scale variations between the donor genome and the reference genome. We used simulation to identify these sources of bias, then quantified these results with measurements on data sets consisting of real reads and sequences from public databases. We have shown that bias can be a significant factor in sequencing experiments.

7.1 Genome Structure

Even with no differences between donor and reference genome, and no errors in the sequencing process, alignment has inherent limitations, because short reads do not always provide enough information to allow unambiguous mapping. Repeats extending beyond the read length are known to create ambiguity (Treangen & Salzberg, 2012), but near-repeats are also problematic. Aligners assign low mapping quality to reads that map to regions that have other similar regions in the reference genome. Any similar region within two differences will result in a low mapping quality, and, if a read contains variations or errors, alternative regions with up to 10 differences affect the mapping quality (see Figure 4.1).

Genomes contain significant levels of repeated sequence, and closely related sequence is also common (see Table 4.1). Levels of repeated sequence vary greatly across genomes, and levels of related sequence are not strongly correlated with levels of exactly repeated sequence. As a result, mappability measures that rely on exactly matching sequence do not capture the effects of closely related sequence, particularly as the effect of near-repeats is magnified in the presence of variation and sequencing errors.

Since lengths of repeated and nearly repeated sequence greatly exceed current read lengths, repeated sequence is expected to be an ongoing limitation of practical sequencing experiments for the foreseeable future.

7.2 Short Variations

When inferring differences between a donor genome and a reference genome, all of the existing limitations present in a variation-free environment apply. Repeats and near-repeats present critical limits to accurate SNV inference. A point mutation introduces a difference between the read and the correct location, which increases the likelihood of alternative ambiguous locations, because any location that differs from the read by one base is now an equally likely candidate for alignment. As more differences accumulate, more distantly related sequences cause ambiguity. The impact of related sequence on accurate SNV prediction is dependent on the level of variation and on the level of repetition in the genome.

Due to the computational infeasibility of a full Smith-Waterman alignment, matching relies on a seed being present – a part of the read that must exactly match the reference. *BWA-MEM's* default seed length is 19 bp and *Bowtie 2's* seed length is 20 bp. If a read does not contain an exactly matching subsequence of this length, the read will not be mapped. With randomly distributed mutations, high prediction accuracy can be achieved at mutation rates up to 4% with the default seed length. Reducing the seed length is an effective method of combating high variation rates: a seed length of 8 bp achieves high prediction accuracy with random mutation rates of 8%.

Similarly, sequencing errors increase the distance between the read and the correct location, but the difference is that, if errors are random, then the likelihood of multiple errors combining to suggest the same variation rapidly diminishes as multiple reads cover a location. At typical mutation and error rates, three reads covering a location significantly reduces the probability of sequencing errors consistently suggesting a variation (see Table 5.1). Consequently, random errors only constitute a problem with low coverage of one or two reads. Assuming a Poisson distribution of reads, we experimentally determined that twenty times sequencing coverage normally provides three times coverage across a genome, which is sufficient for predicting homozygous SNVs that are not in areas of high variability and are not affected by repeated regions. Although repeated regions are a fundamental limitation inherent in the process of aligning short reads, in contrast, high variability is a direct result of the *choice* of reference.

Similarly, indel prediction accuracy is biased by several factors that arise due to the choice of reference. Most indels are affected in a linear, predictable way, which enables errors, clipping, and unmapped reads to be overcome with increased coverage. We calculated the coverage and read length required to recover most indels. However, a small proportion of indels are critically affected by the structure of the surrounding sequence, which greatly reduces their visibility in the alignment. Although many of these problematic regions can be attributed to repetition in the reference genome, some areas are less readily identifiable.

We identified two common structures giving rise to a bias specific to short indels: STRs and similar sequence:

- STRs often result in incorrect alignments, particularly if the length of the indel is a multiple of

the length of the STR, or if the indel can be transformed into the STR for relatively low cost. STRs that extend beyond either end of a read give rise to alignments that are lower cost than the true alignment. This is achieved by effectively moving the indel outside the read.

- Sequence with significant similarity with other regions becomes ambiguous when a variation reduces or removes that difference. This type of bias is particularly likely with longer deletions.

We assessed the capacity of indels to be accurately predicted using the CIGAR string reported by alignment software. We found that, for indels up to 10 bp, over 50% of reads are affected at approximately one location in every 1,000 tested, while 100% of reads are affected at approximately one location in every 10,000 tested. The impact of this is dependent on the sequence coverage, and on whether an affected region is a region of interest. Additionally, variations are more likely in areas of repetition and STRs, which increases the likelihood of a variation being affected by bias.

Two parameters that can improve short variation recovery are sequencing depth and read length. Sequencing depth has limited scope for improvement without increasing read length.

7.3 Whole-Genome Realignment

Aligning a read set to a selection of reference genomes across *E. coli* reveals that the majority of reference bias arises from donor sequence that is missing from the reference. This type of loss is independent of sequencing coverage and largely independent of the read length, and arises purely as a result of the choice of reference genome. Consequently, the more the donor varies from the reference, the less useful the reference genome is as an alignment target. Specifically, if novel sequence appears in the donor that does not appear in the reference, then alignment results are likely to be biased.

Measured reference bias across a selection of *E. coli* strains varies from 0.2% to 29.9%, with a mean of 14.0%, suggesting that reference bias can significantly reduce the ability to accurately reconstruct the donor genome from aligned reads (see Figure 6.5). Although selecting the reference genome with the lowest level of reference bias provides minimal loss of sequence (0.2%), most species do not have a large number of reference genomes available. Even if many potential reference genomes are available, estimating the reference bias inherent in each genome for a specific read set and then selecting the best reference is non-trivial.

This observed bias is a direct consequence of the process of aligning reads to a single reference genome. Although divergent sequence tends to be discarded, novel content is a fundamental limitation of resequencing. Novel content cannot be correctly placed on a reference genome.

7.4 Limitations

We studied a subset of alignment experiments and did not consider a number of adjustments with respect to the types of reads, types of variation, and types of organism. Each of these is addressed below:

- *Paired reads* enable reads in repeated regions to be resolved if one of the pairs is mapped to a unique region. They also provide critical evidence for the presence of SVs. We used long single-ended reads as a proxy for paired reads, but further analysis could more accurately quantify the effect of paired reads relative to single-ended reads.

- *Heterozygous variations* have not been considered. Variations are often expected to be observed in a proportion of the sample. For example, heterozygous variations arise in diploid and polyploidy organisms. Since this increases the difficulty of accurately predicting variations, we expect this to exacerbate the effect of reference bias. However, heterozygous variations were not included in this study.

- *Additional analysis:* we have examined the results generated from an alignment and directly considered the evidence it generates. Reference bias can potentially be reduced by further processing. For example, a significant proportion of reference bias is a result of novel sequence present in the donor that is not present in the reference. We have not assessed the ability of assembly-based techniques to recover and place novel content (Rizk et al., 2014). Similarly, there are numerous solutions that consider additional sources of evidence (Rausch et al., 2012), focus on specific types of variation (Gymrek et al., 2012), or do not use traditional alignment (Abo et al., 2015). These have not been considered.

7.5 Potential Solutions and Further Work

We found that increased coverage improves the ability of variations to be confidently recovered, although there are limits to the benefits due to coverage alone. Increased read lengths improve repeat resolution and enable longer variations to be directly inferred from the alignment.

The alignment step is another source of potential improvements. Aligners require a region of exactly matching sequence in every read, which results in unmapped reads at regions with high variability. When aligning *E. coli* K-12 MG1655 reads to *E. coli* Sakai (see Section 6.2.1), we observed that this requirement was the cause of 40% of unmapped reads. Judicious reduction of the seed length may enable many of these reads to be correctly mapped.

We found that specific combinations of reference sequence and indel can result in particular indels being unusually difficult to find. Extending this analysis could enable mappability estimates to be expanded to include regions where specific variations cannot be predicted.

Ultimately, alignment fails and reference bias arises when there is insufficient unique matching sequence from the donor present in the reference genome. A solution then is to incorporate as much of the donor sequence as possible into the reference genome, by building an ensemble of genomes. This minimises the amount of unrepresented novel sequence from the donor and hence minimises this source of reference bias. Since we determined novel sequence to be a major source of reference bias, this is an approach with significant potential. However, practical implementations are not yet available.

7.6 Conclusion

We have demonstrated that reference bias is a significant problem that can have a considerable impact on resequencing results. Repeats, near-repeats, and STRs cause systematic bias when detecting SNVs and other short variations. This type of bias arises due to the structure of the reference genome and short read lengths. Regions with high levels of variation also cause bias. This bias arises due to aligner limitations. Finally, large-scale variation is a substantial source of bias. The main component of reference bias is caused by novel sequence on the donor that is not present on the reference genome. In *E. coli*, reference bias reached approximately 30% and demonstrates the level of reference bias that can arise when aligning within a species.

References

Abo, R. P., Ducar, M., Garcia, E. P., Thorner, A. R., Rojas-Rudilla, V., Lin, L., et al. (2015). BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Research*, 43(3):e19–e19.

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21(6):974–84, doi:10.1101/gr.114876.110.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195.

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., et al. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Systems*, 1(1):72–87.

Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: accurate indel calls from short-read data. *Genome Research*, 21(6):961–973.

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews*, 12(5):363–376.

Allhoff, M., Schönhuth, A., Martin, M., Costa, I. G., Rahmann, S., & Marschall, T. (2013). Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1.

Babak, T., Deveale, B., Armour, C., Raymond, C., Cleary, M. A., van der Kooy, D., et al. (2008). Global survey of genomic imprinting by transcriptome sequencing. *Current Biology*, 18(22):1735–41, doi:10.1016/j.cub.2008.09.044.

Ball, E. V., Stenson, P. D., Abeysinghe, S. S., Krawczak, M., Cooper, D. N., & Chuzhanova, N. A. (2005). Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human Mutation*, 26(3):205–13, doi:10.1002/humu.20212.

Bao, E., Jiang, T., & Girke, T. (2014). AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics*, 30(12):i319–i328.

Bartenhagen, C. & Dugas, M. (2015). Robust and exact structural variation detection with paired-end and soft-clipped alignments: SoftSV compared with eight algorithms. *Briefings in Bioinformatics*, doi:10.1093/bib/bbv028.

Bashir, A., Volik, S., Collins, C., Bafna, V., & Raphael, B. J. (2008). Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Computational Biology*, 4(4):e1000051, doi:10.1371/journal.pcbi.1000051.

Benjamini, Y. & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10):e72, doi:10.1093/nar/gks001.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., & van Nimwegen, E. (2014). Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Molecular Biology and Evolution*, 31(5):1077–88, doi:10.1093/molbev/msu088.

Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution*, 26(2):473–80, doi:10.1093/molbev/msn275.

Chen, S. F. & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

Cheung, M.-S., Down, T. A., Latorre, I., & Ahringer, J. (2011). Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Research*, 39(15):e103–e103.

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11):987–991.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, 5(6):e11147, doi:10.1371/journal.pone.0011147.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., et al. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8, doi:10.1038/ng.806.

Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., et al. (2012). Fast computation and applications of genome mappability. *PLoS One*, 7(1):e30377, doi:10.1371/journal.pone.0030377.

Dilthey, A., Cox, C. J., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688.

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics*, 14(5):305–314.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, Inc.

Ferragina, P., Luccio, F., & Manzini, G. (2009). Compressing and indexing labeled trees, with applications. *Journal of the ACM*, 57(1):4.

Frith, M. C., Hamada, M., & Horton, P. (2010). Parameters for accurate genome alignment. *BMC Bioinformatics*, 11(1):80.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511.

Genovese, G., Handsaker, R. E., Li, H., Altemose, N., Lindgren, A. M., Chambert, K., et al. (2013). Using population admixture to help complete maps of the human genome. *Nature Genetics*, 45(4):406–14, 414e1–2, doi:10.1038/ng.2565.

Ghanayim, A. & Geiger, D. (2013). Iterative referencing for improving the interpretation of DNA sequence data. Technical report, Technion, Israel, 2013. Retrieved October 1, 2015 from http://www.cs.technion.ac.il/users/wwwb/cgi-bin/tr-get.cgi/2013/CS/CS-2013-05.pdf.

Gontarz, P. M., Berger, J., & Wong, C. F. (2012). SRmapper: a fast and sensitive genome-hashing alignment tool. *Bioinformatics*, 29(3):316–321, doi:10.1093/bioinformatics/bts712.

Gonzalez, K. D., Hill, K. A., Li, K., Li, W., Scaringe, W. A., Wang, J.-C., et al. (2007). Somatic microindels: analysis in mouse soma and comparison with the human germline. *Human Mutation*, 28(1):69–80.

Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: a short tandem repeat profiler for personal genomes. *Genome Research*, 22(6):1154–1162.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, 10(3):R32, doi:10.1186/gb-2009-10-3-r32.

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. V. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14(1):184.

Haubold, B., Klötzl, F., & Pfaffelhuber, P. (2014). andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, doi:10.1093/bioinformatics/btu815.

Hillmer, A. M., Yao, F., Inaki, K., Lee, W. H., Ariyaratne, P. N., Teo, A. S. M., et al. (2011). Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. *Genome Research*, 21(5):665–75, doi:10.1101/gr.113555.110.

Holmes, I. & Durbin, R. (1998). Dynamic programming alignment accuracy. *Journal of Computational Biology*, 5(3):493–504.

Holtgrewe, M., Kuchenbecker, L., & Reinert, K. (2015). Methods for the detection and assembly of novel sequence in high-throughput sequencing data. *Bioinformatics*, doi:10.1093/bioinformatics/btv051.

Homer, N. & Nelson, S. F. (2010). Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biology*, 11(10):R99.

Huang, L., Popic, V., & Batzoglou, S. (2013a). Short read alignment with populations of genomes. *Bioinformatics*, 29(13):i361–i370.

Huang, S., Kao, C.-Y., McMillan, L., & Wang, W. (2013b). Transforming genomes using MOD files with applications. *ACM*, page 595.

Isakov, O. & Shomron, N. (2011). *Deep sequencing data analysis: challenges and solutions*. INTECH Open Access Publisher.

Kayser, M. & de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3):179–192.

Kazazian, Jr, H. H. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664):1626–32, doi:10.1126/science.1089670.

Kehr, B., Trappe, K., Holtgrewe, M., & Reinert, K. (2014). Genome alignment with graph data structures: a comparison. *BMC Bioinformatics*, 15(1):99.

Krawitz, P., Rödelsperger, C., Jäger, M., Jostins, L., Bauer, S., & Robinson, P. N. (2010). Microindel detection in short-read sequence data. *Bioinformatics*, 26(6):722–729.

Lai, D. & Ha, G. (2014). HMMcopy: A package for bias-free copy number estimation and robust CNA detection in tumour samples from WGS HTS data. *R package*.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–9.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–9, doi:10.1038/nmeth.1923.

Langmead, B., Schatz, M. C., Lin, J., Pop, M., & Salzberg, S. L. (2009). Searching for SNPs with cloud computing. *Genome Biology*, 10(11):R134, doi:10.1186/gb-2009-10-11-r134.

Lee, H. & Schatz, M. C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16):2097–2105.

Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 465(7297):473–7, doi:10.1038/nature09004.

Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., & Marth, G. T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One*, 9(3):e90581, doi:10.1371/journal.pone.0090581.

Leinonen, R., Sugawara, H., & Shumway, M. (2010). The sequence read archive. *Nucleic Acids Research*, page gkq1019.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint*, 1303(3997).

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–51, doi:10.1093/bioinformatics/btu356.

Li, H. (2015). BFC: correcting illumina sequencing errors. *Bioinformatics*, page btv290, doi:10.1093/bioinformatics/btv290.

Li, H. & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–95, doi:10.1093/bioinformatics/btp698.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, H. & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–83, doi:10.1093/bib/bbq015.

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–8, doi:10.1101/gr.078212.108.

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., et al. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 333(6038):53–38.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., et al. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19(6):1124–32, doi:10.1101/gr.088013.108.

Li, W., Freudenberg, J., & Miramontes, P. (2014). Diminishing return for increased mappability with longer sequencing reads: implications of the k-mer distributions in the human genome. *BMC Bioinformatics*, 15(1):2.

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108, doi:10.1093/nar/gkt214.

Lunter, G., Rocco, A., Mimouni, N., Heger, A., Caldeira, A., & Hein, J. (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Research*, 18(2):298–309, doi:10.1101/gr.6725608.

Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS ONE*, 9(10):e110808, doi:10.1371/journal.pone.0110808.

Mäkinen, V., Navarro, G., Sirén, J., & Välimäki, N. (2009). Storage and retrieval of individual genomes. In *Research in Computational Molecular Biology*, pages 121–137. Springer.

Mäkinen, V. & Rahkola, J. (2013). Haploid to diploid alignment for variation calling assessment. *BMC Bioinformatics*, 14(Suppl 15):S13.

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11 Suppl):S13–20, doi:10.1038/nmeth.1374.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome Research*, 16(9):1182–90, doi:10.1101/gr.4565806.

Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., Mu, X. J., et al. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Research*, 23(5):749–61, doi:10.1101/gr.148718.112.

Narzisi, G. & Schatz, M. C. (2015). The challenge of small-scale repeats for indel discovery. *Frontiers in Bioengineering and Biotechnology*, 3:8, doi:10.3389/fbioe.2015.00008.

Neuman, J. A., Isakov, O., & Shomron, N. (2013). Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Briefings in Bioinformatics*, 14(1):46–55, doi:10.1093/bib/bbs013.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451.

Osoegawa, K., Mammoser, A. G., Wu, C., Frengen, E., Zeng, C., Catanese, J. J., et al. (2001). A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Research*, 11(3):483–96, doi:10.1101/gr.169601.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., et al. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–78, doi:10.1093/bib/bbs086.

Pearson, C. E., Edamura, K. N., & Cleary, J. D. (2005). Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10):729–742.

Quick, J., Quinlan, A. R., & Loman, N. J. (2014). A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. *GigaScience*, 3(1):22.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, doi:10.1093/bioinformatics/bts378.

Reinert, K., Langmead, B., Weese, D., & Evers, D. J. (2015). Alignment of next-generation sequencing reads. *Annual Review of Genomics and Human Genetics*, 16:133–151, doi:10.1146/annurev-genom-090413-025358.

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, 72(4):686–727, doi:10.1128/MMBR.00011-08.

Rizk, G., Gouin, A., Chikhi, R., & Lemaitre, C. (2014). MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–7, doi:10.1093/bioinformatics/btu545.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., et al. (2015). The UCSC genome browser database: 2015 update. *Nucleic Acids Research*, 43(Database issue):D670–81, doi:10.1093/nar/gku1177.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51.

Ruffalo, M., LaFramboise, T., & Koyuturk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796.

Satya, R. V., Zavaljevski, N., & Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic Acids Research*, 40(16):e127, doi:10.1093/nar/gks425.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., & Gibrat, J.-F. (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology*, 19(6):796–813, doi:10.1089/cmb.2012.0022.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–5, doi:10.1126/science.1178534.

Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O., et al. (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98, doi:10.1186/gb-2009-10-9-r98.

Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., et al. (2011). Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences*, 108(25):10249–10254.

Schrider, D. R., Gout, J.-F., & Hahn, M. W. (2011). Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One*, 6(10), doi:10.1371/journal.pone.0025842.

Schröder, J., Girirajan, S., Papenfuss, A. T., & Medvedev, P. (2015). Improving the power of structural variation detection by augmenting the reference. *PLoS One*, 10(8):e0136771, doi:10.1371/journal.pone.0136771.

Schröder, J., Hsu, A., Boyle, S. E., Macintyre, G., Cmero, M., Tothill, R. W., et al. (2014). Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8):1064–1072, doi:10.1093/bioinformatics/btt767.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1):16–8, doi:10.1038/nmeth1156.

Schwartz, S., Oren, R., & Ast, G. (2011). Detection and removal of biases in the analysis of next-generation sequencing reads. *PLoS One*, 6(1):e16685, doi:10.1371/journal.pone.0016685.

Sharp, A. J., Locke, D. P., McGrath, S. D., Cheng, Z., Bailey, J. A., Vallente, R. U., et al. (2005). Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, 77(1):78–88, doi:10.1086/431652.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

Siren, J., Valimaki, N., & Mäkinen, V. (2011). Indexing finite language representation of population genotypes. *Algorithms in Bioinformatics*, pages 270–281.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., & Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*, 21(10):1728–37, doi:10.1101/gr.119784.110.

Slater, G. S. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–7.

Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. *BMC Genomics*, 14(1):536.

Tang, Y.-C. & Amon, A. (2013). Gene copy-number alterations: a cost-benefit analysis. *Cell*, 152(3):394–405.

Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S., & Salim, A. (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718.

The 1000 Genomes Project Consortium et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Trapnell, C. & Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nature Biotechnology*, 27(5):455–457.

Treangen, T. J. & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, doi:10.1038/nrg3117.

Vellai, T. & Vida, G. (1999). The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1428):1571–1577.

Wetterstrand, K. (2015). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). Retrieved November 1, 2015 from http://www.genome.gov/sequencingcosts.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10:451–81, doi:10.1146/annurev.genom.9.081307.164217.

Appendix A

Data

This appendix describes all data sets used in this thesis.

A.1 Individual Genomes

- Drosophila melanogaster chromosome 2L (Ensembl Release 81) – 23 513 712 bp.
- Escherichia coli O157:H7 str. Sakai (NC 002695.1) – 5 498 450 bp.
- Escherichia coli str. K-12 substr. MG1655 (NC 000913.3) – 4 641 652 bp.
- Homo sapiens chromosome 6 (GRCh37) – 171 115 067 bp.
- Homo sapiens chromosome 21 (GRCh37) – 48 129 895 bp.
- Human immunodeficiency virus 2 (HIV-2) (NC 001722.1) – 10 359 bp.
- Plasmodium falciparum (3D7 version 13) – 23 332 831 bp.

A.2 E. coli Phylogeny

The accession codes of the 62 *E. coli* reference genomes used to study bias across a phylogeny are available at: https://github.com/supernifty/mgsa/blob/master/mgsa/data/ecolis.txt.

A.3 Read Sets

- Escherichia coli str. K-12 substr. MG1655 (SRR892241) – 7.1M paired reads of length 200 bp.

Appendix B

Software

All custom code used to perform simulations and experiments is available at http://github.com/supernifty/mgsa/. In addition, we used the following software and versions:

- Andi 0.9.4 (Haubold et al., 2014)
- Bedtools v2.23.0-24-g7553f4a (Quinlan & Hall, 2010)
- BFC r181 (Li, 2015)
- Bowtie 2 (Langmead & Salzberg, 2012)
- BWA-MEM 0.7.12 r1039 (Li & Durbin, 2010)
- Exonerate 2.2.0 (Slater & Birney, 2005)
- GATK v3.3-0-g37228af (DePristo et al., 2011)
- IGV 2.3.59 (Robinson et al., 2011)
- Progressive Mauve 2015-02-13 (Darling et al., 2010)
- R 3.2.1 (R Core Team, 2015)
- SAMtools 1.2-2-gf8a6274 (Li et al., 2009a)

Glossary

allele-specific expression The situation where the two alleles of a gene are expressed at different rates.

chromosome A single piece of DNA.

CIGAR A string produced by alignment tools to describe how a read maps to a reference, including matches, insertions, deletions and substitutions.

copy number variations Large-scale duplication of DNA.

de novo assembly Assembling a genome from reads without relying on a reference genome as a template.

deoxyribonucleic acid The molecule carrying genetic instructions enabling a cell to operate.

diploid A cell that has two copies of each chromosome.

eukaryote Organism consisting of cells with a nucleus.

genome The genetic material of an organism.

haploid A cell that has a single copy of each chromosome.

homopolymer Sequence that consists of the same repeated base.

indel An insertion into or deletion from a DNA sequence.

k-mer A distinct substring of a string of length k.

mapping quality Represents the aligner's confidence in the correctness of a particular read's alignment.

paralog Paralogs are genes related by duplication within a genome, with potentially differing function.

phenotype The observable characteristics or traits of an organism.

polyploid A cell that has more than two copies of each chromosome.

prokaryote Single celled organism that lacks a nucleus.

pseudogene A copy of a gene that has mutated to become non-functional.

reverse complement In DNA, A and T are exchanged, G and C are exchanged, and the order is reversed.

short read Sequence extracted from a DNA fragment, whose length ranges from tens to hundreds of base pairs.

short tandem repeat Directly adjacent repeating sequence, usually with period less than 10 bp. Also known as microsatellites.

short variation A difference that is shorter than the read length so can potentially be directly inferred from the aligned reads.

shotgun sequencing A sequencing technique that breaks a DNA sequence up into smaller fragments.

single-nucleotide polymorphism Substitution at a single base that is considered to be "common" across a population. "Common" often is taken to mean greater than 1%.

single-nucleotide variation Substitution at a single base that is not considered to be common across a population.

structural variation Large-scale non-local variations to a genome.

transcription The process of transcribing a DNA subsequence to RNA.

translation The process of translating RNA to a protein.

transposable element mobile DNA sequences that can move or copy themselves to different regions.