Analysis of short tandem repeat variation in large scale resequencing data

Weldon W. Whitener

Wellcome Trust Sanger Institute

St. Edmunds College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

September 2011

I would like to dedicate this thesis to my loving parents...

Acknowledgements

None of this would of been possible without the love, support and guidance from the best parents a son could ever hope to have. Thank you Mom and Dad for always believing in me.

Now, where to begin?

There have been so many people who have contributed in making this document possible. Along the way, I've had more support than anyone could ask for, as well as, plenty of lucky breaks! I want to begin by thanking all my professors back at my *alma mater*, North Carolina State University. Specifically, I would like to thank Dr. Frank Abrams for always allowing me to use the 'big words', as well as, Dr. Elizabeth Loboa for giving me my first shot at scientific exploration.

I want to thank Dr. Mark van Dyke for allowing a scrappy biomedical engineer into one of the most amazing labs I've ever had the good fortune of being a part of. You are one of the most inspiring and friendly persons I've ever met and I am forever grateful for your support and guidance along the way. None of my subsequent success would of been possible had I not met you by chance at a career event my third year at North Carolina State University.

The next person I'd like to thank is someone whose guidance and support cannot be fully conveyed in writing. Thank you Jennie LaMonte for always going above and beyond. More than anything, you have been a great friend and mentor to me for the past six years and

I hope that we remain friends for many years to come.

The past four years have been some of the most challenging, yet thought provoking, years of my life. Richard Durbin accepted me into his group and helped me cultivate a probing and scientific outlook that has helped me answer some of the most difficult questions I've ever been posed. Richard's continued support, guidance and occasional knock on the head has given me a true confidence in myself that I will be forever grateful for. Thank you for teaching me to stand on my own two feet and giving me the tools to help me be successful no matter where life takes me.

This thesis would not of been possible without the help of countless people at the Sanger Institute and University of Cambridge. Thank you Leopold Parts and Aylwyn Scally for helping me come to grips with statistical modeling. You were the best labmates I could of asked for and the help you offered me will never be forgotten. Thank you Jim Stalker, Thomas Keane and the entire vertebrate sequencing group for making my life so much easier by doing a great job at curating all the sequencing data – an almost insurmountable task! Thank you Avril Coghlan for your continued collaboration throughout the duration of my thesis. Thank you Stijn van Dongen, Sergei Manakov, David Adams, Theo Whipp and Steve Russen for your support and friendship throughout my time at the Sanger Institute. I also wish to send a huge thank you to my favorite collaborator, David Knowles. Without David's masterful grasp of machine learning and diligence, most of the higher level analysis conducted in the latter chapters of this thesis would not of been possible. David even managed to make it a fun process – through his energy and good nature – which by no means is no small feat!

Lastly, I want to thank a person who, as much as my parents, has helped me become the the person I am today. Without my sister

Witnee's guidance, I would not of been able to achieve all that I have. She taught me how to conduct myself in a manner that has opened doors for me both socially and professionally. She also offered the best advice while keeping my spirits high. I could not of done this without her, and for that, I am forever grateful.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text. This thesis does not exceed the length limit set by the Biology Degree Committee.

Weldon W. Whitener
30 September 2011

Abstract

The average eukaryotic genome contains many types of variation; from single nucleotide polymorphisms, small, medium and large insertions and deletions to copy number variation, translocations and inversions to name a few. The genome is also highly non-uniform, with some regions more variable than others. Tandem repeats are stretches of DNA comprised of a short motif repeated end-to-end multiple times. They are of interests to geneticists because they exhibit a high rate of length variation and are relatively frequent in the genome. However, until now they have been hard to assay using new sequencing technologies, which have revolutionized the study of other types of genetic variation. In this thesis, we address this deficit by developing methods to genotype short tandem repeats from shotgun short sequencing reads and applying them to human genome data.

To begin, I present a statistical model based on a Bayesian framework which uses Illumina paired end sequencing reads to determine the genotype of a diploid individual at a given short tandem repeat locus. This method is applied to all triplet tandem repeats (repeat motifs three bases in length) in the human genome for an individual sequenced deeply from multiple libraries as part of the 1000 Genomes project. We show that our method has good sensitivity and specificity for both homozygous and heterozygous indel genotypes measuring over three bp in length.

Next, we build upon the previous chapter by utilizing our model for genotyping across nine deeply sequenced individuals. We use the putative indel calls made in this data set to gain an understanding of

what factors of a tandem repeat have the largest effect on observing an indel at a given locus. We look at the effect that various measures of repeat length, repeat purity, GC content and tandem repeat motif have on triplet repeat variation. This analysis furthers our understanding of tandem repeat variation.

Lastly, we reformulate our individual genotyping model to take sequencing data from multiple, low sequence depth individuals in a population to understand the population distributions of variants at tandem repeat loci. This uses machine learning approaches including the expectation-maximization algorithm and Gibbs sampler, that help elucidate which loci show evidence of variation in the sample population, and allow us to explore the distribution of alternate alleles at a locus. As well as cataloguing variation efficiently, this allows us to examine a broader picture of the contribution the previously described factors have in influencing variation at a tandem repeat locus.

Contents

List of Figures

Chapter 1

Introduction

Revolutions in science have often been preceded by revolutions in measurement.

– Sinan Aral, a business professor at New York University (The Economist [2010]).

1.1 New age of technology

The age of modern technology has led to a paradigm shift in regards to how scientific exploration is conducted. Where once data collection limited our ability to answer pressing questions about highly complex systems, we are now capable of generating far greater amounts of data at a fraction of the time and cost. As the capacity of digital devices increase while the price decreases, the amount of information we are now privy to is magnitudes in size larger than before. Simply, the amount of digital information increases approximately tenfold every five years while Moore's law states that processing power and storage capacity of computer chips double (or their prices halve) roughly every 18 months (Moore et al. [1998]) which in turn drives our current accumulation of data. However, along with all the benefits of this data comes the problem of how we make inference about the underlying systems at play.

With magnitudes more data at hand, it has become an important goal of science to develop algorithms and models which can make sense of all this new information. When utilized to their full potential, large data sets can provide

fresh insights into many natural systems. The intrinsic make up of many of these systems lend themselves perfectly to a highly computational and statistical approach: from analysing high energy physics data to forecasting weather. While each system has its own intricacies, the prevailing concepts on the underlying mechanisms are closely related to one another such that advancements in one field can benefit another field's exploration (Cohen [2004]). One system which has enjoyed many advancements through both direct design and from crossover synergies is DNA sequencing. Where it once took ten years for the first few human genomes to be sequenced (International Human Genome Sequencing Consortium [2001], Levy et al. [2007]), the time frame has been lowed to approximately a single week to sequence an entire human individual's genome. The per base cost of DNA sequencing has lowered to about 100,000x cheaper than it was a decade ago (Nature Jobs [2011]). This abundance of data has increased the need of computational approaches, algorithms and statistical models to make new discoveries which rely less on the biochemistry of the system and more on the complexities that arise from such large data sets. Given the raw data from DNA sequencing, geneticists have endeavored to develop algorithms and models which can reveal new insight into the complexities of the genome that would previously have remained hidden. This new world of genomic sequencing has given credence to the belief that genomic medicine has a bright future once geneticists and bioinformaticians decipher the context of the genome. It is only a matter of time before the "base pairs to bedside" concept is a reality (Green et al. [2011]).

1.2 Sequencing technology and bioinformaticians

The emergence of new sequencing platforms has chauffeured in a new type of geneticist: a scientist with proficiency in both computer science and statistical theory who is able to disambiguate the needle of truth from the haystack of data. The paradigm shift from benchtop to laptop has changed the way genetic research is conducted. The need for these newly trained scientists far outstrips the current supply which necessitates the migration of individuals into this field (Nature Jobs [2011]). However, the need for quantitatively trained geneticists hasn't always been the case in the field of sequencing whose history stretches back over

four decades.

As with all technological movements, sequencing has experienced a number of periods that are described by the technology and knowledge of the time. Starting with the sequencing of RNA by Frederick Sanger (Brownlee et al. [1967]) and the subsequent sequencing of DNA (Sanger et al. [1982]), this process has been an archetypal example of exponential technology growth. After Sanger sequencing came high throughput DNA sequencing that was conducted using electrophoretic methods in miniaturized systems; such as capillaries, capillary arrays, and microchannels (Carrilho [2000]). We are now in what is known as the the next generation sequencing era which is comprised of a number of platforms, processes and chemistries (Metzker [2009]). These new sequencing technologies have effected a change within genetics; one where the sequencing of a full genome to a reasonable depth is no longer prohibitively expensive. The speed and low cost has led to a number of resequencing projects aimed at demarcating variants within multiple species' genomes.

1.3 Genomic variation

Single nucleotide polymorphism (SNPs) represent the largest class of variation within the human genome, but a large number of 'structural variations' have been uncovered as well. Small insertions and deletions (indels) represent the second most frequent class of variation in the human genome followed by deletions, duplications, inversions, translocations and other large-scale copy-number variants. An important class of indels within short tandem repeats or microsatellites (characterized by having multiple exact or near exact tandem copies of a 1-20 bp sequence motif) will be the main subject of this thesis which we will return to later. While indels exhibit a greater potential to disrupt functional elements compared to SNPs, they have been characterized to a lesser extent. Because of this, they are under represented in public variation databases; while there are 24,359,333 unique SNPS in the dbSNP database (version 132), there are only 5,617,945 short indels. Furthermore, resequencing projects have also shown that structural variants can comprise megabases of nucleotide heterogenity within a

given genome and are likely to make an important contribution to human diversity as well as disease susceptibility (Feuk et al. [2006]).

1.4 Detecting small scale insertions and deletions

Whole genome sequencing using next generation sequencing technologies has shown that several hundred thousand indels are located in a single individual's genome compared to the reference genome (Wheeler et al. [2008]; Bentley et al. [2008]; Wang et al. [2008]; McKernan et al. [2009]). Various methods have been proposed in locating these sites with the most common being based on the aligning of sequenced reads directly to the reference and searching for specific signals that are indicative of a breakpoint. This can be accomplished directly by the split alignment (or gapped alignment) of reads which span across a breakpoint. Essentially, if a read from a sequenced individual contains inserted or deleted sequence relative to the reference sequence, the read will not map exactly to the genome. Reads whose prefix and suffix match a specific region in the reference to some identity can then either have sequence removed – with the ends appended to one another (deletions) – or be split at some distance in the reference (insertions) to determine if the read then matches the reference genome. Variations of this approach have been used by numerous sequence alignment algorithms (Li et al. [2008]; Homer et al. [2009]; Li and Durbin [2009]; Rumble et al. [2009]) which have located many of these small indels within a resequenced genome. This is not a perfect method, however. Reads that span a break point close to its end have been shown to be difficult to align and can lead to misalignment and in turn false SNP calls (Krawitz et al. [2010]). This problem has been mitigated through the local realignment of reads which span a putative break point (McKenna et al. [2010]; Homer et al. [2009]; Albers et al. [2011]). Further, many of these tools do not permit gaps above a certain size in their split alignments. The maximum gap size is due in part to the computational cost it would require to search for larger

and larger gaps and in part because allowing larger gaps can lead to errors. Depending on the algorithm, the cost to search possible gap sizes grows non-linearly. Some aligners will use the Smith-Waterman algorithm to map reads which on the first pass are not mapped correctly to the genome. Most aligners allow user input to dictate the aggressiveness of resolving gaps. These values can be tweaked to allow larger gaps, but run the risk of having more false indel discoveries. However, if the deletion is too large, then the flanking sections will be shorter and there will be too many places within the genome the two end lengths of a read (split by a deletion) can be placed. Similarly, the size of detectable insertions is only a few base pairs, as every inserted base reduces the fraction of the read that matches the genome (Medvedev et al. [2009]). Because of this, most indels of more than a few bases in size are not detected by standard split alignment methods.

A few methods, such as PolyScan, have been developed to locate short indels of size $\leq$ 100 bp by analysing long reads from capillary sequence data (Chen et al. [2007]). As with the previously mentioned alignment tools, PolyScan aligns reads to the reference genome and infers indels from gaps in the alignments. This can be used to infer indels in many of the unique regions of the genome. However, as well as the size of the indel, the efficacy of calling indels is contingent upon the reads being mapped uniquely to the reference genome. In unique regions of the genome this is not a problem, but as the uniqueness of DNA decreases, so does an aligner's ability to map a read correctly to a specific position on the reference genome. Nowhere is this more problematic than in repetitive copies of DNA which take various forms within a genome. Copy number variation (or CNV) represents the largest type of repeating patterns where whole regions of DNA are duplicated throughout the genome. Mapping to these regions is difficult as it is usually unknown which copy the sequenced read is coming from.

1.5 Tandem Repeats

A particular form of repeat region that is prevalent in the genome and contains length variation that is hard to type is tandem repeats (minisatellites). These regions are characterized by 21-60 bp repeat units that are repeated in a tandem end-to-end fashion some number of times in the genome consisting of both full or truncated repeat patterns as well as pure and impure repeat tracts. The smaller equivalent of tandem repeats – and the more prevalent form – are known as short tandem repeats (or microsatellites). Short tandem repeats (STRs) are repetitive segments of DNA that are characterized by 1-20 bp repeat units. As with tandem repeats, they can be both full or truncated repeat units consisting of both pure and impure repeat tracts. Altogether, there are over 2.1 million STR loci of motif lengths 1-10,15 and 20 located in the human reference genome.

The STR sites were located by running Tandem Repeats Finder (TRF) version 4.00 (Benson [1999]) across the entire human reference genome (NCBI build 36). TRF is able to locate both pure and impure (interrupted) repeats using a probabilistic model of tandem repeats. Essentially, TRF aligns two tandem repeat copies of some motif pattern of length n by a sequence of n independent Bernoulli trials. A Bernoulli trial is defined as a number of independent repeated trials of an experiment with only one of two outcomes: success or failure (or match and mismatch in our case). The probabilities of these outcomes are then defined as p for the probability of success and $q = 1 - p$ for the probability of failure. A series of Bernoulli trials which consists of n trials is known as a binomial experiment. The probability of k success out of n trials can then be written as

$$P(k) = \binom{n}{k} p^k q^{n-k}$$

For TRF's purposes, the probability of a base matching the pattern (success), $P(match)$, is representative of the average percent identity between copies. For mismatches (SNPS), insertions or deletions, a second probability is described, $P(mismatch)$. This denotes the average percentage of mismatches, insertions and deletions between the copies. TRF uses the distribution of the Bernoulli se-

quences to locate tandem repeats within the genome to some stringency defined by the properties of the alignment ($P(match)$ and $P(mismatch)$). These bounds, $P(match)$ and $P(mismatch)$, serve as a type of extremal limit – a quantitative description of the most divergent copies TRF will report.

TRF is broken down into two components: detection and analysis. The program first locates candidate regions in the genome which can be described as tandem repeats and then the analysis component attempts to produce an alignment at each of the candidate sites and if successful, produces a number of statistics about the alignment and sequence (percent identity, percent indels, composition and entropy measure).

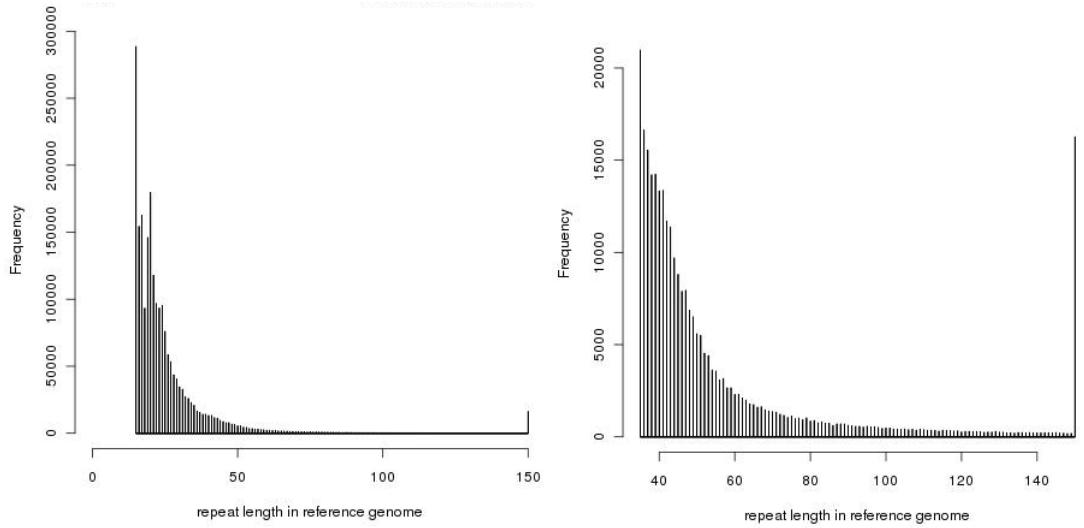
The detection step is broken down into a series of algorithms which scan through the genome looking for repetitive patterns known as $k-tuples$. A $k-tuple$ is a window of k consecutive characters from a nucleotide sequence. Matching $k-tuples$ are two windows with identical contents and if aligned in the Bernoulli model would produce a run of k successes. Once these sites are identified, the candidate pattern corresponding to some positions in the genome are selected from the nucleotide sequence and aligned with adjacent sequence. If at least two copies of this pattern are aligned correctly, the tandem repeat is reported. After these patterns are matched, an initial candidate pattern P is drawn from the sequence. TRF then iterates through possible patterns from the sequence until a consensus pattern by majority rule is found from the alignment of P copies back to the candidate region. This consensus sequence is then used to realign the sequence and the final alignment is reported with the respective period size of the repeat motif.

TRF uses a number of parameters which the user can define in regards to the stringency of locating tandem repeats within a genome. The parameters correspond to the alignment weights for match, mismatch and indels, the matching probability and indel probability, a maximum period size for patterns to report and a minimum alignment score to report a tandem repeat. In our analysis, we left most parameters in the out-of-box configuration. We did, however, iterate

through each repeat motif length we were interested in looking at. We also set the minimum alignment score to report a repeat to 30, which corresponded to a 15 bp perfect triplet repeat or a longer impure triplet repeat. In addition to this criteria, all repeats (independent of their motif length) were required to be at least 15 bp in length. In total, TRF identified 2,136,510 repeats in the human reference genome that met this criteria. This amounted to over 58 Mb of genomic sequence in the human genome. The results of our TRF run are summarized in table 1.1 and figure 1.1.

Loci, base count and statistics for STRs in the human genome				
Motif size	Loci count	Bases	Mean	Std Dev
1	447847	9705850	21.672	6.684
2	209248	7655889	36.588	45.909
3	86401	2391275	27.676	41.335
4	267055	9232626	34.572	53.215
5	168674	4892872	29.008	240.971
6	218574	4949601	22.645	22.739
7	291167	5910812	20.300	29.382
8	207127	4986481	24.075	26.304
9	151583	4067068	26.831	85.026
10	39215	1505968	38.403	56.680
15	28833	1533692	53.192	94.687
20	20786	1533188	73.761	121.431
total	2136510	58365322	27.318	102.788

Table 1.1: Counts of all tandem repeat loci found by TRF within the human reference genome that correspond to a given repeat motif with corresponding mean and standard deviation statistics. The first column represents the motif size and the second and third column represents the number of loci and total bases, respectively, corresponding with the motif length in the human genome. The fourth and fifth columns are the calculated mean and standard deviations for all loci in that row, respectively.



(a) Histogram of lengths of STR loci in human genome

(b) Histogram of lengths of STR loci greater than 35 bp in the human genome

Figure 1.1: Histogram of number of loci of each length across the human genome with loci longer than 150 bp binned in the last bin. The number of STR loci (motifs of 1-10, 15 and 20 bp) across the genome are mostly of lengths <40 bp (1,926,168 of 2,136,510, roughly 90%). Even the shortest paired end reads (36 bp) are almost able to extend across these repeat loci to make indel calls by alignment possible (given the indel is not an insertion that increases the repeat length above the length of the short paired end read). This limits the amount of sites which our model is applicable (see table 1.3) for high coverage data sets. However, samples sequenced with paired end reads at a lower coverage will have much lower chance of reads being sequenced exactly so that they can expand across an STR locus (see figure 1.4).

Aside from their prevalence in the human genome, STRs come in a variety of lengths within the genome. While the average length of STRs is around 27 bp, the standard deviation is extremely large as shown in table 1.1. This large discrepancy in the sizes of the standard deviations – specifically for motifs of lengths 5 and 9 bp – are most readily explained by extremely long loci. While most motifs' longest loci are anywhere from two to six thousand bp in length, the motifs of lengths 5 and 9 bp have loci that are as long as sixty-five and twenty-five thousand bps in length, respectively (see table 1.2).

Ten longest loci for each motif length	
Motif size	Ten longest loci
1	92, 93, 97, 98, 99, 101, 113, 128, 396, 415
2	1620, 1636, 1645, 1710, 1740, 1741, 1801, 1838, 1844, 4760
3	1314, 1321, 1354, 1509, 1528, 1722, 1804, 2594, 3148, 3925
4	2027, 2093, 2162, 2173, 2531, 2963, 3144, 4101, 5656, 6240
5	4863, 4927, 6585, 7433, 26557, 26771, 28286, 29067, 46493, 65350
6	1383, 1428, 1436, 1509, 1537, 1589, 1780, 1826, 1835, 2403
7	1989, 1996, 2045, 2065, 2067, 2295, 2339, 2365, 3024, 4816
8	1328, 1357, 1494, 1497, 1577, 1613, 1835, 1919, 2180, 2779
9	2331, 2531, 2892, 3783, 3861, 4107, 5651, 6235, 10241, 25733
10	1348, 1358, 1414, 1504, 1527, 1632, 2086, 2182, 2229, 2266
15	2305, 2309, 2366, 2403, 2590, 2713, 2830, 2837, 2865, 4327
20	2032, 2205, 2362, 2432, 2533, 2555, 2600, 2784, 4139, 4360

Table 1.2: Lengths of the ten longest loci in each tandem repeat length motif.

1.6 Small scale insertions and deletions in tandem repeats

1.6.1 Background significance of tandem repeat indels

While also being extremely prevelant in the human genome, tandem repeat loci are highly variable between populations and individuals due to their relatively high mutation rate compared to the rest of the genome (Pearson et al. [2005]). They commonly undergo indel mutations of single or multiple repeat units (Di Rienzo et al. [1994]), thus the two copies of a locus in an individual may easily differ by up to 100 bp from that in the reference genome. Small indels have been shown to be more prevalent in tandem repeat regions of exons than in non-tandem repeat regions of exons. Tandem repeat loci that lie within exons have been shown to be significantly over-represented in disease-related genes in both human and mouse (Madsen et al. [2008]). Indels in both coding and non-coding tandem repeat loci have been linked to diseases such as spinocerebellar ataxia (SCA types 1, 2, 3, 6, 7), Huntingtons disease, fragile X syndrome, and myotonic dystrophy (Ball et al. [2005]; Hamosh et al. [2005]). To date, tandem repeat instability has been implicated as the causative factor in more than forty

neurological, neurogenerative and neuromuscular disorders (Pearson et al. [2005]) by pathogenic mechanisms involving the loss or gain of function at the protein or RNA level (Gatchel and Zoghbi [2005]). While tandem repeat loci of all repeat unit sizes are susceptible to mutations, triplet repeats have come to the forefront of tandem repeat research due to the high number of diseases caused by indels at triplet repeat loci (Pearson et al. [2005]). We note that triplet repeats are relatively rarer in the sequence than other short motif tandem repeats (see table 1.1) and wonder whether it is possible that this is due to some form of selection.

Tandem repeat loci evolve mainly through replication slippage-mediated gain and loss of single repeat units (Ellegren [2000]; Mahtani and Willard [1993]). Recent studies have shown that, in addition to replication slippage, expansions and contractions at tandem repeat loci can also be caused by faulty repair of DNA lesions (Kovtun and McMurray [2008]; Lenzmeier and Freudenreich [2000]). Given their abundance and high mutation rates, tandem repeat loci play an important role in the ongoing evolution of the human genome (Ellegren [2004]). It is very likely that some indels in tandem repeat loci are the cause of normal phenotypic variations in humans and other species (Kashi et al. [1997]; Kashi and King [2006]). In addition to their importance to disease and evolution, variation at tandem repeat loci has been very useful in ascertaining the demographic history of human populations throughout the world (Zhivotovsky et al. [2003]).

1.6.2 Detection of indels using paired end mapping information

Carrying on from table 1.1, it is important to keep the distribution of tandem repeat lengths in mind when we start to look at calling indels within a tandem repeat. Indels in repeat regions can be called in a similar way as indels within unique regions of the genome. However, directly calling indels within tandem repeats from split alignments only works up to a point. When the total length of the repeat in the sequenced individual increases towards the read length, the read can no longer be aligned accurately to the reference genome. Reads whose sequence is comprised entirely of a repeating pattern are unable to be mapped

correctly to the genome for multiple reasons. One such instance is when a read is sequenced from a CNV because it is difficult to tell which copy the sequenced read is coming from. Similarly, as tandem repeats are the same pattern of sequence repeated over and over, there is no way of telling which of the many STR loci in the genome with the same motif a read is sequenced from, nor where in the repeat locus the sequenced read should be placed. This causes a problem when trying to determine the exact length of a tandem repeat locus, and in turn, whether a sequenced individual contains an insertion or deletion. One way to rectify this problem has been to target sequence these loci with longer reads, for example from capillary sequencing. Another way has been to target a specific locus by PCR with primers in flanking unique sequence, but this is low through put by modern standards. The large amount of money and time needed to genotype many tandem repeat loci has been prohibitively expensive and because of this, typing these sites on a large scale has been difficult. However, the chemistry for some next generation sequencing technologies provides additional information that can be used to solve this problem: the sequenced reads are paired, which correspond to two regions that lie some genomic distance apart in the genome of the sequenced individual. This distance (or fragment length) is a consequence of the sizes of DNA fragments selected by virtue of coming from the two ends of a DNA fragment created during library construction. Read pairs that are proximal to the tandem repeat on each side of it but not within the repeat locus are mapped to the reference genome and the additional mapping distance data offers information in determining the length of the tandem repeat. Therefore, instead of a read being 36 bp in length (a standard read length for early sequencing from the Illumina platform), the physical coverage (or distance between mapped reads) increases the pair's reach up to hundreds of base pairs that can now span across a repetitive region and offer information about the repeat tract's length in a sequenced individual. It is through this paradigm that many of the next generation indel callers identify longer indels.

As alluded to in section 1.4, extensive sequencing of tandem repeat loci has been limited due to the costs and time required using traditional capillary sequencing

methods. Compared to traditional capillary sequencing methods, next generation sequencing machines produce orders of magnitude more sequence data in a fraction of the time and cost (Mardis [2008]). The trade-off is that the sequenced reads for platforms, such as Illumina, are much shorter than traditional capillary sequence reads – currently around 100 bp in length per read for the Illumina platform. From these shorter reads, multiple tools have arisen to fill in the gap left by alignment tools to find indels larger than a few base pairs.

The concept of using end sequencing profiling (ESP), also known as paired end mapping (PEM), of paired reads to demarcate structural variations has been around since 2005. Applied to both somatic structural variations in cancer genomes (Volik et al. [2006]) and normal genomes (Tuzun et al. [2005]), what these methods have in common is that they use the distribution of the distance between the paired end reads to facilitate researchers' ability to locate large insertions and deletions. Essentially, these algorithms assess the distribution of paired end read separations mapped to a reference genome and define cutoffs where they feel the mapped separation of two reads in the reference was more extreme then expected, and occurred because of a structural variant rather than by chance. The earlier incantations of this methodology used fosmid pairs to locate very large insertions and deletions by locating regions in the genome where the paired alignment of reads mapped anomalously. These algorithms looked for reads which mapped further than three standard deviations away from the mean (Volik et al. [2006], Tuzun et al. [2005]), and at a certainty of over 99%, these 'discordant' reads (reads whose mapping was not in line with the distribution) were indicative of a structural variant. When these discordant pairs occurred in clusters at a specific genomic region, they gave more power to make a putative variant call (see figure 1.2). However, as the fosmids' separations were so large, the resolution to find variants was limited to structural variants on the order of tens of kilobases and larger. As technology evolved, this methodology migrated over to next generation sequencing technologies – such as Korbel's use of the 454 platform (Korbel et al. [2007]). As fragments from next generation sequencing machines were smaller and in turn more tightly distributed, the resolution to find smaller variants became possible. In line with previous studies, Korbel defined a

cutoff distance for paired end reads which was indicative of a variant. Through this method, variants of size 2 kb and larger were located in the human genome.

As the methods of fragment library creation become better, the distribution of fragments became tighter and so did the ability to call smaller and smaller indels. Using the more recent sequencing of both the 454 and Illumina platforms, structural variation callers can be broken down conveniently into three subgroups – with each subgroup having its own process of locating indels of varying sizes.

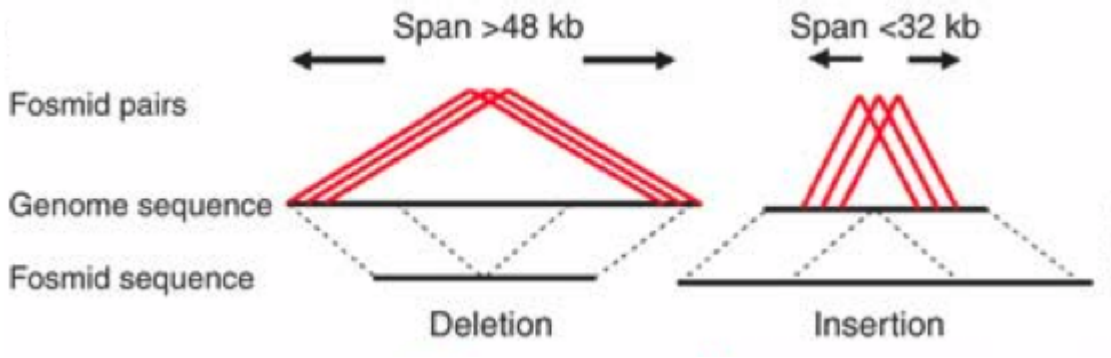


Figure 1.2: The underlying paired end sequencing methodology used to detect structural variation by fosmid pairing (Tuzun et al. [2005]). Deletions in the fosmid source are defined as sites where two or more fosmid end-sequence pairs span $>$ 48 kb. Insertions are defined as sites where two or more fosmids span $<$ 32 kb (red). These length thresholds are three standard deviations from the mean insert size.

The first, and smallest group, is comprised of the Geometric Analysis of Structural Variants tool (GASV). This algorithm takes a geometric approach for structural variation identification, classification and comparison. Instead of using the paired read separations directly to locate discordant reads and then make inference, this approach represents the uncertainty in the measurement of a structural variant as a polygon in the plane and identifies measurements supporting the same variant by computing intersections of polygons (Sindi et al. [2009]). This work was the first of its kind to present a general framework for comparing structural variants across multiple samples and measurement techniques. While this paper presented a very interesting way to think of structural variants, the methods where not used extensively within the field of structural variation detection. The previous paradigm of finding outliers remained the prevailing technique for locating structural variations.

The next group of callers can be seen as a direct extension of Korbel, Tuzun and Volik's outlier methods. First, extending further on his research, Korbel released PEMer (or paired end mapper) in 2009 (Korbel et al. [2009]). Using the same strategy as in his first paper, Korbel looked for clusters of various read numbers to locate discordant reads whose separations were greater than three standard de-

viations away from the median. However, unlike his previous method, PEMer's methods were applied across multiple sequencing platforms: 454, Illumina, and ABI. The efficacy of PEMer's modeling was tested on the 454 platform and had very marginal gains in being able to detect smaller indels than previously listed. Also, in the same year, two more tools were released which boasted a higher resolution for calling smaller indels using the same principle of looking for clusters of reads mapping some number of standard deviations away from the mean. As well as PEMer, SVDetect also used multiple sequencing platforms to locate large, genomic structural variations (Zeitouni et al. [2010]), but lacked the power to call significantly smaller indels. This was was answered by two other structural variation callers: VariationHunter and McKernan's SOLiD method. VariationHunter (Hormozdiari et al. [2009]) was able to locate deletions and insertions smaller than 100 bp using Illumina paired end reads as the libraries were much tighter than that of the 454 platform. The paired end reads used for this analysis came from a single individual having a sequence depth of roughly 42x and a physical coverage of 120x (fragment size of 200 bp, Bentley et al. [2008]). Next, McKernan published a paper using the SOLiD platform to locate deletions as small as 86 bp and insertions as small as 30 bp. As the sizes of indels being found reached their maximum resolution given the current technology and methods, it was necessary to re-evaluate the method which only looked for discordant reads which mapped some number of standard deviations away from the mean/median.

In the same year as many of these other tools came out, two algorithms came out which took a novel approach to calling indels: BreakDancer and MoDIL. BreakDancer, like many of the other tools, used discordant reads whose mapped separation was outside three standard deviations to locate structural variants. Using this method, it was run on a data set consisting of 844 structural variants identified on chromosome 17 of J. Craig Venter's genome: 425 deletions, 415 insertions and 4 inversions ranging from 20 to 7,953 bp. Paired end reads were simulated measuring 50 bp in read length at 100x physical coverage with a normally distributed insert size library with a mean size of 200 bp and standard deviation of 20 bp. While able to locate many variants at a decent sensitivity, 38.4% (324 including 147 shorter than 60 bp), and a low false positive rate,1.48%,

it had trouble locating the smaller indels as well as variants which occurred in repetitive regions that are difficult to map to or assemble across. In addition to this, the novel part of BreakDancer included an additional method – named BreakDanceMini – designed to locate smaller indels in the region of 10 to 20 bp. Instead of only locating the regions of discordant reads mapping largely away from the mean, it took anomalous regions (areas where a cluster of reads where larger than expected but less so than discordant reads) and compared the distributions of the paired end mappings of these regions with the full data set of paired end separations using a two-sample Kolmogorov-Smirnov test. If the K-S statistic measured ≥ 2.3 (indicating the distribution of separations are in fact different) the locus was tagged as a variant. The use of the Kolmogorov-Smirnov test increased the number of false positives to 10%, but also increased the method's ability to call 10-20 bp indels.

Before moving on to the last tool, I will provide a bit of background on the Kolmogorov-Smirnov test (K-S test). The K-S test is a nonparametric test for the equality of continuous, one-dimensional probability distributions that is used to compare both a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test). These two tests quantify the distance between the empirical distribution function of the sample and the cumulative distribution of the reference distribution or the distance between the empirical functions of the two samples. The null hypothesis for these two tests is that the sample is drawn from the null distribution (one-sample) or that both samples are drawn from the same distribution (two-sample). Essentially, the K-S test can serve as a goodness of fit test between multiple distributions. The empirical distribution function F_n of n *independent identically distributed* (*iid*) observations X_i is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}$$

where $\mathbb{I}_{X_i \leq x}$ is the indicator function (equal to 1 if $X_i \leq x$ and equal to 0 otherwise). For clarity, *iid* – as referred to previously – is a term in probability theory

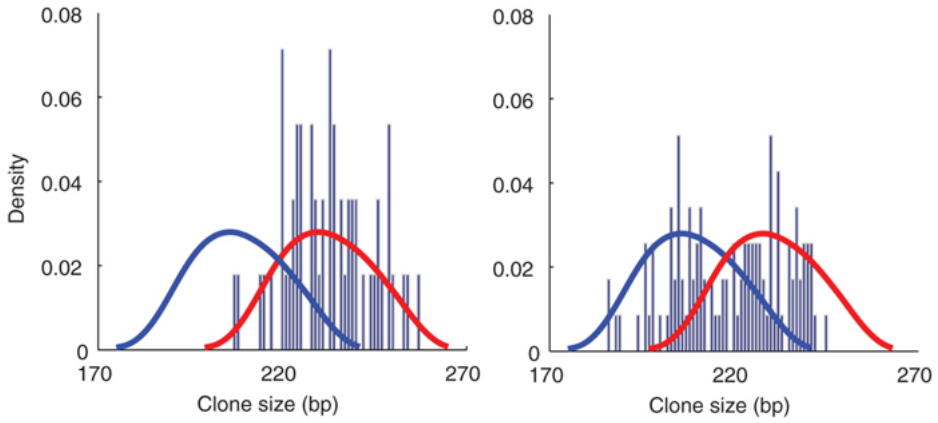
and statistics that defines a sequence – or other collection of random variables – that each random variable has the same probability distribution as the others and are mutually independent. From this, we are able to define the K-S statistic for a given cumulative distribution function $F(x)$ as

$$D_n = \sup_x |F_n(x) - F(x)|$$

where $\sup_x$ is the supremum of the set of distances, and if the sample comes from the distribution $F(x)$, then D_n converges to 0 almost surely with increasing n. In analysis, supremum (or least upper bound) of a set S of real numbers is defined to be the smallest real number that is greater than or equal to every number in S. A critical value of D_n is set such that any time the test statistic is above the critical value, the null distribution is rejected – that the sample distribution was not drawn from the null distribution. This knowledge is important when describing the methods of the MoDIL tool.

MoDIL (mixture of distributions indel locator) was the first method to specifically look for indels in the size range of 20 to 50 bp from next generation sequencing data. As with BreakDancerMini, MoDIL is not limited in resolution of structural variation detection by searching only for large paired end read deviations, but uses clustered reads whose deviation by a small number of nucleotides is indicative of an insertion or deletion. The MoDIL algorithm, instead of looking for discordant read pairs, compares the distribution of paired end separations in the sequenced library to the distribution of observed paired end distances at a particular genomic location. By streaming through the genome, MoDIL looks at each genomic location and clusters paired end reads which overlap a particular position. At sites where there is no indel, the distribution of paired end separations at a genome location should match the distribution of all paired end separations across the genome. However, if there has been a homozygous indel at this location, the distribution will shift off the population distribution by approximately the size of the indel. If there is a heterozygous indel, there will then be two distributions from which the paired end separations will come from with approximately half of the paired end reads coming from one distribution and half

from the other (see figure 1.3). MoDIL represents the genotype of a putative



(a) Distribution of paired end mappings for a homozygous deletion.

(b) Distribution of paired end mappings for a heterozygous genotype

Figure 1.3: Example of a homozygous (1.3a) and heterozygous (1.3b) deletion with the observed distribution of mapped distances shown in gray. (1.3a) A homozygous deletion of 24 bp. Notice the shift from the null distribution (blue) to the best match distribution (red). (1.3b) A heterozygous deletion of 24 bp with one allele the same length as the reference length. The mapped distances at this locus are generated from two distributions with means centering at 230 bp and 208 bp (deletion and reference allele, respectively).

variant locus by the random variable of the expected size of the indel (the mean of the fragment library size minus the paired end read separation) with two random variables representing each haplotype. From each cluster, MoDIL tried to identify the two distributions, {D1,D2}, with fixed shapes and arbitrary means that best fit the observed data using the K-S test. When locating the means of the two distributions, MoDIL employs an expectation-maximation algorithm with appropriate Bayesian priors to prevent over-fitting. By assuming that the reads are drawn from a single fragment library with a defined distribution which follows a Gaussian distribution with some known mean and standard deviation, MoDIL iterates through possible genotypes and reports which indel pair value minimizes the goodness of fit test from the K-S test.

MoDIL has shown promise in locating and describing smaller indels than the

previously described tools. By looking at smaller variations in the paired end mappings rather than very large divergences, it has been able to locate much smaller indels within the genome. However, MoDIL is weakened in the long run by some of the assumptions it makes. These assumptions are that the distribution of paired end separations is well defined by a Gaussian distribution and that all the reads come from a single distribution. While the aim of fragment library creation is to have a tight, well described distribution of paired end separations, this is not always the case. Also, individuals are often sequenced by multiple fragment libraries. Because of this, a hole exists in the current literature on how to address the mass sequencing now being undertaken at sequencing centres across the world. Lastly, none of these tools – including MoDIL – are specifically designed for typing tandem repeat regions. None of the aforementioned tools take into consideration some of the bias that occurs in paired end read mappings around tandem repeats which unchecked, could lead to many false positives. As discussed earlier, read mapping to tandem repeats becomes more and more difficult as the repeat length increases.

Split alignments are only able to call extremely short indels (a few bp in length) in short repeats, while paired end mapping tools are unable to accurately and consistently call small indels (5-20 bp). This leaves an important part of genomic variation un-assayed on a large scale, as shown in table 1.3. More importantly in the case of split alignments, a read must to not only span the repeat, but also extend a sufficient distance into the proximal unique sequence on each side to place it unequivocally at this particular repeat in the genome.

1.6.3 Relevance of an indel caller for short tandem repeats

In determining the necessity of developing an indel caller specifically for tandem repeats, we looked at whether the previous gapped alignment tools were sufficient enough to answer this problem. In doing so, we calculated the expected number of times a region (or repeat tract) would be both extended across by reads of a given length as well as physically covered (spanned). Reads of length 100 bp, as well as fragment libraries of size 300 and 500 bp, were chosen as they are most

Loci lengths for various motif lengths					
Motif size	Total	≥40 bp	≥60 bp	≥80 bp	≥100 bp
1	447847	9930 (2.217%)	770 (0.172%)	74 (0.017%)	5 (0.001%)
2	209248	55765 (26.650%)	14896 (7.119%)	8453 (4.040%)	5821 (2.782%)
3	86401	8295 (9.601%)	2806 (3.248%)	1892 (2.190%)	1385 (1.603%)
4	267055	46166 (17.287%)	23612 (8.842%)	15859 (5.938%)	11712 (4.386%)
5	168674	17709 (10.499%)	6117 (3.627%)	3150 (1.87%)	1977 (1.172%)
6	218574	10562 (4.832%)	3498 (1.600%)	1767 (0.808%)	1075 (0.492%)
7	291167	5443 (1.869%)	1955 (0.671%)	1314 (0.451%)	1009 (0.347%)
8	207127	9116 (4.401%)	3894 (1.880%)	2468 (1.192%)	1751 (0.845%)
9	151583	6429 (4.241%)	2777 (1.832%)	1816 (1.198%)	1337 (0.882%)
10	39215	8537 (21.770%)	3985 (10.162%)	2388 (6.090%)	1672 (4.264%)
15	28833	12067 (41.851%)	3681 (12.767%)	2069 (7.176%)	1419 (4.921%)
20	20786	20323 (97.773%)	6073 (29.217%)	3165 (15.227%)	2150 (10.344%)
totals	2136510	210342 (9.845%)	74064 (3.467%)	44415 (2.079%)	31313 (1.466%)

Table 1.3: Count of tandem repeat loci of lengths for a given motif repeat length. The second column shows the number of loci in the human genome of that given motif length. The third through sixth columns are the number of loci (and percent of total) of greater than or equal length of that in the header of the column (lengths 40, 60, 80 and 100 bp). The shorter read lengths of most new sequencing technologies means that many loci would remain un-assayed by split alignment methods.

representative of what is currently being sequenced by the Illumina platform. The coverage (c) and number of reads (z) were the most important factors to take into consideration as they are essential in determining the expected number of extending and spanning reads for the aforementioned scenarios.

Base pair coverage and physical coverage are calculated in the same way, the difference being the length of the segment (b) and number of reads; the single ends will consist of two times more reads than the paired ends as each pair is comprised of two single end reads. Coverage, can generally be calculated the same way as in equation 2.7 (described in detail in chapter 2. Conversely, being interested in the number of reads, a simple reorganization of equation 2.7 yields the number of reads produced at a given coverage

$$z = \frac{c \cdot g}{b}$$

where depending on if you are looking for single or paired end reads you may keep or omit the coefficient two in the denominator, respectively. Next, we calculate the number of subregions (s_q) of a given length (q) that are within the entire region we are sequencing. This will aid us in determining how often each of these subregions are extended/spanned across by our single and paired end reads

$$s_q = g - q + 1$$

We next calculate how many subregions (t_q) are crossed by each of the single and paired end reads for a given q. This can be calculated identically as the number of mappable positions, p_m, was in equation 2.1 (see chatper 2 for further discussion). Lastly, we can directly calculate the expected number (f_q) of times a subregion is extended/spanned across by single and pair ended reads

$$f_q = \frac{z \cdot t_q}{s_q}$$

Figure 1.4 illustrates the expected number of extending and spanning reads you would observe for a given coverage across STRs of varying size.

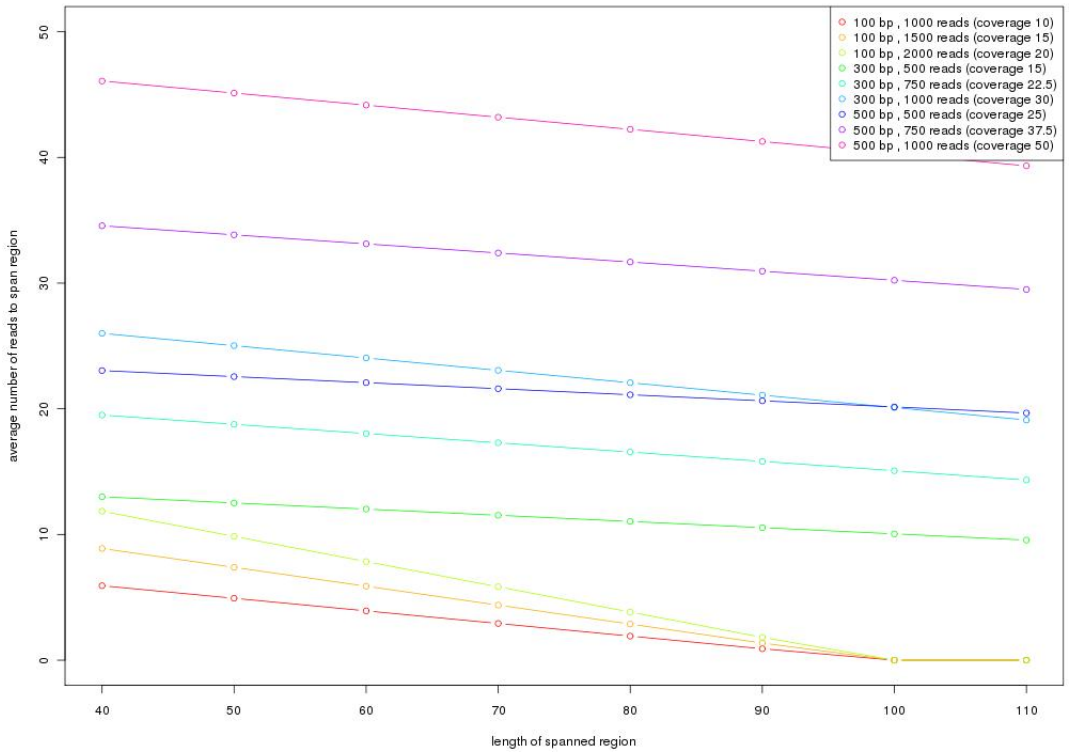


Figure 1.4: Graph of expected number of spanning reads (physical coverage) and reads that extend across various genomic lengths at base pair coverages of 10, 15 and 20x. Reads of length of 100 bp were chosen to illustrate the upper bound of read lengths currently available. The spanning coverage was then calculated for fragment libraries of sizes 300 and 500 bp. It is clear from the graph that although many sites will have a few extending reads, all sites will have multiple spanning reads which can be used to ascertain whether an indel exists in a given repeat tract. Most callers-by-alignment need at least 2 to 3 reads to extend across a region to make an accurate call as there is a chance that a singleton may be a read sequencing error – especially in repeat tracts. This means that the cutoff for being able to make calls using crossing reads is lower than the read length.

1.7 Ascertaining tandem repeat allele frequencies in large populations

High throughput sequencing technologies have made population scale sequencing studies of genetic variation a reality. The 1000 Genomes Project has been one

of the most recent large scale population sequencing projects to come out of the next generation sequencing era. It has aimed to provide a deep characterization of human genome sequence variation as a foundation for understanding the relationship between genotype and phenotype. As low frequency variants (those defined as having a minor allele frequency between 0.5 and 5%) vastly outnumber common variants, and are also believed to contribute significantly to disease susceptibility, it was the goal of the 1000 Genomes Project to systematically locate these variants across the global population to facilitate further research and our understanding of how genetic diversity contributes to phenotypic expression. Overall, the project aims to characterize over 95% of variants that are in genomic regions accessible to current high throughput sequencing technologies that have an allele frequency of at least 1%.

The 1000 Genomes Project's design is to sequence populations in each of five major continental groups (ancestry in Europe, East Asia, South Asia, Africa and the Americas) to an average depth of 4x. In the recent low-coverage sequencing pilot study, 179 individuals were sequenced to roughly 2-6x using a mix of platforms, with about 80% of reads coming from the Illumina sequencers. In total 60, 59 and 60 individuals were sequenced from the CEU, YRI and CHB+JPT populations with a collective total number of mapped bases at 1,881 Gb (3.56x coverage). The current Phase 1 build of the 1000 Genomes Project has over 1000 individuals sequenced from 14 populations (see figure 1.5). From the pilot sequencing, researchers were able to identify 14.4 millions SNPs, 1.3 million short indels and over 20,000 larger structural variants. The FDR for this set was experimentally validated to be kept below 5% for SNPs and short indels, and less than 10% for structural variants. This pilot study has shown the power, and in turn efficacy, of pooling individuals together in similar populations to demarcate variation. Understanding genome variation is well within scientists' grasp and it is only a matter of time before all variation to a very low frequency will be found. However, the one caveat to many of these large sequencing projects is the amount of inaccessible regions that arise from the low coverage and short read lengths. Of the reference genome, 85% was readily accessible in the 1000 Genomes Pilot project as well as 93% of the coding sequences. Of the 15% that remains

inaccessible, 97% has been annotated as repeats or segmental duplications. Repeats remain an area of low penetrance for calling both SNPs and indels. The

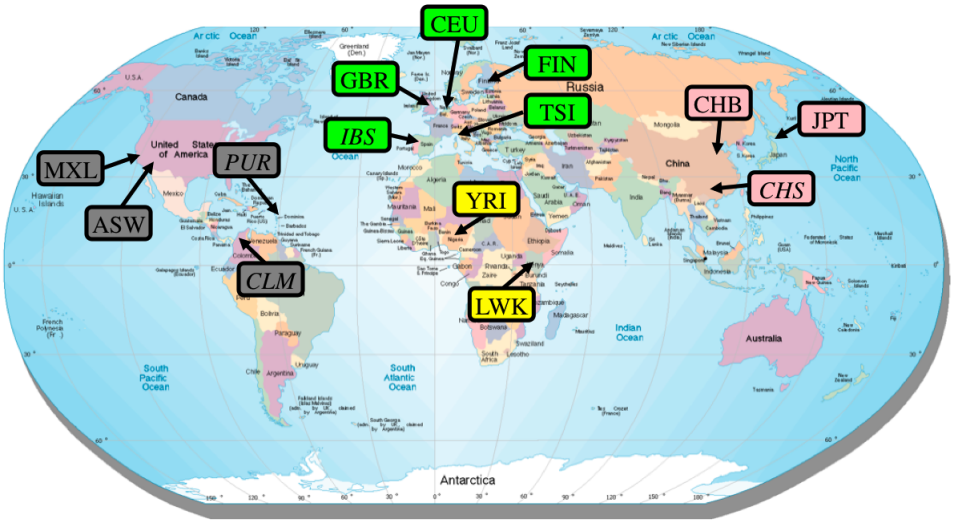


Figure 1.5: Map of populations in 1000 Genomes Project Phase 1 build.

sheer number of individuals sequenced in many of these studies limit the effective coverage by which each individual can be sequenced to. This in turn can make calling certain variants a difficult task. Past population sequencing projects using capillary technology (Bhangale et al. [2005]; Mills et al. [2006]) have elucidated some variation on a population scale, but the inherent cost of sequencing large parts of the genome using the Sanger method has proved prohibitively expensive for a full genome assay of indels.

Alongside the 1000 Genomes Project, methods for demarcating variation in pooled populations has been a large point of research over the past few years. Some models have been developed which aim at finding the actual genotype of each individual within a population by using the background population sequencing as a context from which an individual's reads are compared (Bansal and Libiger [2011]). Essentially, they propose that the evidence supporting a variant allele at a position in an individual will be significant when compared to the population background in the absence of that variant. A likelihood ratio test is used to compare the results of an individual's sequencing to that of the population

where a cutoff is put in place so that any individual's loci that are above this cutoff are assigned the putative genotype. In total, 408 indels were identified across seven populations in the 1000 Genomes exon sequencing data. As these regions were sequenced to a high depth by both 454 and Illumina sequencing, the promise of this method locating many indels across the entire genome is quite low.

Another suggested approach is using the pooled information to learn the shared variation amongst a population rather than solely use the population as a background parameter against which to compare an individual's data. This comes from the knowledge that each read corresponds to a specific allele length in an individual that is also part of the overall allele frequency in the population. These reads can therefore be leveraged with one another to accurately detect variant frequencies within a population. This has allowed population geneticists to identify both common and rare DNA sequence variants within a population (Koboldt et al. [2009]). These methods have previously been developed for SNPs, but no such methods have been developed to specifically look at highly polymorphic tandem repeat loci.

1.8 Proposal

In chapter 2 of this thesis, I present a novel method that uses the additional read mapping information to analyse Illumina sequencing data to probabilistically model the length of the two copies of a tandem repeat locus in a sequenced individual. This method will allow me to genotype any deep sequenced individual at any short tandem repeat locus whose repeat length is below the fragment library length. This method is then applied in chapter 3 to nine deeply sequenced individuals. The resulting genotypes of these individuals at each locus will be combined and used in understanding what increases the probability of observing a variant at a short tandem repeat locus. In chapter 4, I reformulate my genotype calling method for low coverage individuals who are sequenced as part of large resequencing projects – such at the 1000 Genomes Project. This population variation method will use the combined information from sequenced reads in all

individuals in a population. This population based approach intends to understand the underlying distribution of variants at a locus within a population. This model can be used to explore what sites are actively evolving and what sites' allele distribution is not best explained by the reference.

Chapter 2

Genotyping short tandem repeats using short paired end reads from two deeply sequenced individuals

Collaboration note *This chapter contains work performed in collaboration with Dr. Avril Coghlan. Avril assisted in the identification of tandem repeats in the human genome using Tandem Repeat Finder, as well as designing and implementing a method for determining the haplotype of multiple sequenced individuals using trace reads from the Trace Archive (Cochrane et al. [2009]) which was instrumental to determining a prior probability of observing an indel of a given magnitude.*

The largest hindrance in genotyping a STR locus arises as the repeat length approaches, and ultimately surpasses, the length of a read. This makes it extremely difficult for assemblers as they are unable to accurately determine the exact placement of a read within the locus as there is no point of reference. Some assemblers will estimate the repeat length based on the coverage of reads in the repeat locus (Myers [2005]). This assumption, however, is highly variable as the effective read coverage across the genome is subject to random fluctuations, and even when the read depth is very deep, it is not consistent (Bentley et al. [2008]) yielding inaccurate length predictions.

However, due to the advent of paired end read sequencing, we now possess additional information that can be used in determining the length of a tandem repeat by modeling the expected separation of the two reads. This process, as it turns out, is not as straight forward as one might imagine, as there are many considerations that must be taken into account when modeling the expected separation of the reads in a sequenced pair.

2.1 Locating tandem repeats in the human reference genome

We began our analysis of STRs by first locating all tandem repeat positions in the human genome. We relied on Tandem Repeats Finder (TRF) version 4.00 (Benson [1999]) to locate all repeat loci in the human reference genome (NCBI build 36) corresponding to repeat motif lengths of 1-10, 15 and 20 bp. TRF was able to locate both pure and impure (interrupted) repeats. The minimum alignment score to report a repeat was set to 30, which corresponded to a 15 bp perfect triplet repeat or a longer impure triplet repeat. In addition to this criterion, all repeats (independent of their motif length) were required to be of at least 15 bp long. In total, TRF identified 2,137,399 repeats in the human reference genome that met this criteria. The results of our TRF run are summarized in table 1.1 in chapter 1 (which represents the number of loci after migrating the positions from NCBI build 36 to GRCh build 37, described below).

2.1.1 Translating NCBI build 36 coordinate to GRCh build 37 coordinates

Over the course of this project, it was necessary to migrate the tandem repeat coordinates from NCBI build 36 to GRCh build 37 as newer sequence runs' reads were mapped to GRCh build 37 and older reads were remapped to the newer coordinates. LiftOver (Kuhn et al. [2006]) was used as it was able to realign the tandem repeat positions to the newer coordinates from a chain file which was downloaded from

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/
Almost all of the positions were able to be migrated uniquely, though due to changes in the reference, 889 sites were not used due to being partially or fully deleted, or split in the newer GRCh build 37 genome for all tandem repeat lengths (1-10, 15 and 20 bp).

Looking at the triplet repeats, 86,435 loci were identified in NCBI build 36 with 86,401 uniquely migrated to GRCh build 37 (34 excluded loci: 7 deleted and 27 partially deleted). A by eye analysis of these loci using the UCSC Genome Browser (http://genome.ucsc.edu) showed liftOver's results to be correct; that these loci had in fact been removed or relocated somewhere up or down stream in GRCh build 37. As the number of sites unable to be accurately migrated over was deemed insignificant (0.04% for triplet repeats), we did not feel it was necessary to rerun TRF on the new GRCh build 37 genome.

2.2 Sources of sequence

Two sequenced individuals were used for this project; NA12878 and NA18507. Both samples were sequenced on the Illumina platform which generates paired end reads from the two ends of DNA fragments that were size selected during library creation. In addition to the Illumina sequence, NA12878 was also sequenced on the 454 platform. Both individuals had some additional shotgun genome sequence obtained using traditional Sanger (capillary) methods. These additional sequences were indispensable to the modeling and validation of our method. The long capillary reads were necessary to help establish the prior parameters for our model and also served as an *ad hoc* resource in locating candidate sites for validation by 454 reads.

2.2.1 Individual NA12878 sequence

The sequence data for both the Illumina and 454 platforms are available from ftp://ftp-trace.ncbi.nih.gov/1000genomes/.

2.2.1.1 Illumina read sequence data

As part of the pilot project of the 1000 Genomes Project (Consortium [2010]), individual NA12878 (the daughter of a HapMap father-mother-daughter trio of European ancestry) was sequenced to approximately 22.5x sequence depth with paired end reads of average read length 37 bp on the Illumina platform.

2.2.1.2 454 sequence data

In addition to Illumina sequencing and as part of the pilot project of the 1000 Genomes Project, individual NA12878 was sequenced to approximately 12.8x sequence depth with an average read length of 276 by the 454 platform.

2.2.1.3 Capillary sequence data

We downloaded 2,156,700 reads pertaining to individual NA12878 from the ERA trace archive with an average read length of 722 bp and at an average depth of coverage of 0.5x.

2.2.2 Individual NA18507 sequence

2.2.2.1 Illumina read sequence data

The genome of a male Yoruban individual, NA18507, was fully sequenced by the Illumina sequencing platform (Bentley et al. [2008]) to an average depth of 41x sequence coverage with paired-end reads, whose average read length length was 32 bp. The Illumina sequence data for NA18507 is publicly available in the short read archive by accession SRA000271
(http://www.ncbi.nlm.nih.gov/sra/SRA000271).

2.2.2.2 Capillary sequence data

We downloaded 3,916,150 reads pertaining to individual NA18507 from the ERA trace archive with an average read length of 741 bp and an average depth of coverage of 0.9x.

2.3 Mapping of paired end reads to the human reference genome

Each sequenced individual's short paired end reads were aligned to the reference human. Reads from individual NA18507 were aligned with MAQ (Li et al. [2008]) to NCBI build 36. Reads from individual NA12878 were aligned using BWA (Li and Durbin [2009]) to GRCh build 37 along with other 1000 Genomes samples.

When working with paired end reads' mapping data, it was necessary to acquaint ourselves with the various mapping scenarios one would encounter. As the focus of our analysis is on tandem repeats, I will limit the type of mapped paired end read scenarios to the following (though this is not exhaustive and ignores unmapped paired end reads as well as reads which would signify inversions and translocations, (Korbel et al. [2007])): uniquely mapped paired end reads, spanning paired end read pairs and hanging/anchoring reads. By far the largest group are uniquely mapped paired end reads, which as their name states, are mapped uniquely anywhere within the genome and are constrained only by their mapping quality (described in 2.4). The group of reads that will be the focus of this chapter are spanning paired end reads. Last are hanging/anchored reads which arise around repeats due to the inability of a read to map uniquely through a repeat as seen in figure 2.1.

2.4 Determining the empirical distribution of a given library's mapped paired end read separations (MPERS), P(M)

One of the principal factors in determining the genotype of a STR locus using read pair data is first knowing the distribution of separations for a given library. The distribution of lengths of the DNA fragments from which paired end reads were sequenced can be estimated by mapping all reads to the reference genome and calculating the distance between the mapped positions of the two reads of each read pair (the mapped paired end read separation, MPERS, see section 2.3

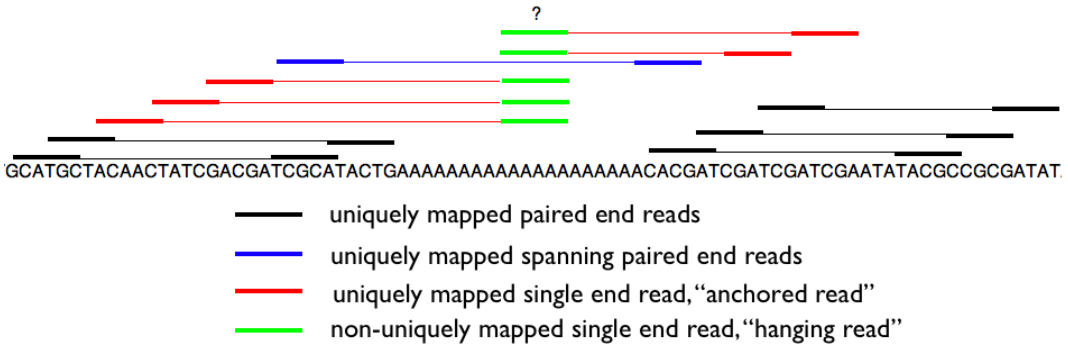


Figure 2.1: Four of the various mapping scenarios related to paired end reads. Paired end reads which map uniquely within the genome and are filtered only by their mapping quality are known as unique reads (black). Paired end reads which are of sufficient length and have mapped on either side of a repetitive region are known as spanning reads (blue). Adjacent to repetitive regions lie anchoring reads which map to the unique flanking regions of a repeat (red) and whose mate (green) maps within the repetitive region.

for read mapping). The MPERS distribution is different for each sequencing library, because each library is in general made from a different preparation of DNA fragments.

We were able to calculate the MPERS distribution for each library quite simply. After alignment of the sequenced reads to the reference genome, it was only a matter of parsing through the alignment file and applying the following calculation: if the first read of a read pair mapped to coordinates $x_1 - x_2$ on a chromosome in the reference genome and the second read mapped to coordinates $x_3 - x_4$ on the same chromosome on the reference genome (where $x_2 > x_1$, $x_4 > x_3$ and $x_3 \geq x_1$), the MPERS (M) is the distance between the start of the mapped position of the first read (x_1) and the end of the mapped position of the second read (x_4) plus 1; $M = x_4 - x_1 + 1$.

The empirical distribution of MPERS for all read pairs from each library was calculated from approximately ten million uniquely mapped paired end read pairs. We refer to the empirical distribution of MPERS for all read pairs from a library as $P(M)$. This is an estimate of the probability distribution of the lengths of

the fragments in the library. Thus, the mean of the $P(M)$ distribution is an estimate of the mean size of the fragments in that library. Often, an individual was sequenced from multiple fragment libraries and therefore had multiple MPERS distributions.

After mapping the paired end reads to the reference genome (see section 2.3), we were left with alignment files detailing the mapping position of each paired end read to the chromosome to which it was mapped. Starting with chromosome 1, we streamed through the alignment files taking only paired end reads whose single ended mapping quality score, q, was equal to or above 30 (this corresponds to a mapping error rate of $\leq$0.001 as taken from PHRED scoring (Ewing and Green [1998]) where error $= 10^{-q/10}$). We believed it was important for our analysis that both reads mapped uniquely to the reference. It is not unusual for the paired end mapping score to be much higher than the single ended mapping score and this is never more the case than when looking at repetitive regions in the genome. The discrepancy between single ended and paired end mapping scores arises due to the fact that the paired end mapping score makes use of the additional information of what the expected paired end mapping separation should be. This is a problem for our calculation when one of the reads maps to a unique position while the other maps into non-unique sequence. While the read that is mapped to the non-unique sequence is unable to be placed exactly, the knowledge from its mate limits the range by which it is placed. This causes the paired end score to be much higher than the single ended score. This is a major problem for our model when we rely on the exact mapping of both reads to determine the MPERS. By limiting our assessment to only mate pairs that are made up of two reads that both map uniquely independent of one another, we were able to remove any systematic bias that might occur both in a library's MPERS distribution as well as our actual genotype predictions (described below in section 2.6.3.1). It was also important that the two reads be mapped in the correct orientation with respect to one another. Incorrect orientations could signify an inversion or translocation (Korbel et al. [2007]) which would only act to obfuscate our model and predictions and are outside the scope of this analysis.

2.4.1 The empirical distribution of mapped paired separations (MPERS)

2.4.1.1 Individual NA12878

Individual NA12878 was sequenced from eight separate paired end read libraries. Of these eight libraries, two were not considered in our analysis as none of their paired end reads mapping qualities were above our set PHRED score of 30. The six libraries used in our analysis varied in genome coverage from 1.5 to 6.4x. Because of the lower depth sequencing of some libraries, we were unable to locate ten million uniquely mapped pairs for every library. We simply took as many reads as we could find and from them, generated the empirical distribution of each library. The statistics for each library are seen below in table 2.1.

Library statistics for individual NA12878				
Library	Bases sequenced	Mean	STD	Coverage
g1k-sc-NA12878-WG-1	19327027164	301.1	144.6	6.4
Solexa-3630	14717717437	83.8	9.1	4.9
g1k-sc-NA12878-CEU-1	12546297144	140.9	12.5	4.2
NA12878.1	10463534460	232.4	11.0	3.5
g1k-sc-NA12878-CEU-2	6012622836	180.7	31.0	2.0
Solexa-5460	4443002700	204.9	31.4	1.5
totals	67510201741	196.3	52.2	22.5

Table 2.1: Statistics for individual NA12878's libraries. Columns (from left to right) represent the library name, the number of sequenced bases, the mean value of the MPERS, the standard deviation of the MPERS and the overall base coverage in the genome.

2.4.1.2 Individual NA18507

Individual NA18507 was sequenced from a single short paired end read library from which we calculated the MPERS for ten million uniquely mapped paired end read pairs. These read pairs had a near Normal distribution of MPERS ranging from 36-270 bp, a mean MPERS of 209 bp and a standard deviation of 13 bp ($\sim$6.2% of the mean). The shortest observed MPERS of 36 bp would arise when

each of the reads in a read pair mapped to exactly overlapping positions in the reference genome.

2.5 Detecting indels in tandem repeat loci using long capillary reads from the Trace Archive

We detected indels in tandem repeats by analysing aligned traditional (capillary) sequence reads downloaded from the Trace Archive. For our analysis, we only considered repeat loci that have unique flanking regions to ensure that reads matching a locus were not from a paralogous locus. Repeat loci with unique flanking regions were verified using SSAHA2 (Ning et al. [2001]) by searching for matches in the reference genome to the sequence 100 bp up and downstream of each tandem repeat site. A 100 bp flanking region was considered unique if it only had a match to itself, or if its best non-self match had <90% identity.

At each tandem repeat locus with two unique flanking sequences, we used the Trace Archive SSAHA2 Client (Ning et al. [2004]) to search for matches between its 100 bp flanking sequences and human reads in the Trace Archive. A read matching the flanking regions of a tandem repeat locus was accepted if: (i) it had matches of ≥97% identity to both flanking regions and the matches were in the same order as in the reference genome; (ii) the matches covered ≥80% of both flanking regions; and (iii) the repeat locus in the read had high quality sequence (all bases had PHRED (Ewing et al. [1998]) quality scores of >10).

Indels in tandem repeats were then identified by finding cases where the length of a repeat locus differed between the reference genome and a matching sequence read from the Trace Archive. To estimate the length difference, the read was aligned using SSEARCH (Pearson [1991]) to a sequence consisting of the reference genome repeat locus plus 100 bp of up and downstream DNA. The length of the gapped region (if any) in the repeat locus in the SSEARCH alignment was used as an estimate of the length difference between the reference genome's length and sequenced sample's length.

Of the matches between the capillary reads and tandem repeats, many contained an identifier for the individual from whom the DNA originated. As the coverage was quite low, we were only able to determine one haplotype at most individuals' loci, but in a few cases we had evidence that led us to believe we could correctly genotype an individual at a given locus, that is, determine both haplotypes. This was only possible if we detected two distinct alleles at a tandem repeat locus using the Trace Archive reads from an individual. We therefore assumed that the individual must be a heterozygote at that locus and therefore knew the true genotype. On the other hand, if we only detected one allele at a particular repeat locus using Trace Archive reads from an individual, it was impossible for us to know whether the individual is homozygous at the locus or heterozygous with only one allele represented in sequenced reads in the Trace Archive. Due to the random nature of shotgun sequencing (Anderson [1981]), by chance some sites were sequenced more than others. Sites which contained more spanning traces gave us more information in regards to whether the site truly was homozygous. For instance, looking solely at traces which contained a unique identifier for triplet repeat positions in the human genome (219,796), the Trace Archive contained 3,654 individuals' positions which contained at least 4 spanning reads. Knowing that there is a 50% probability being drawn from one allele or the other, the probability of observing (or not observing) one of the alleles can be described by the binomial distribution. For the case of observing a reference allele in four traces, the probability of observing only the reference allele in a heterozygote by chance is 6.25%. This knowledge becomes important when considering which sites were best suited for validation (2.9.2.1). An initial set of 3,534 trace calls from individual NA18507 was used to generate the prior probability distribution for a single allele call in our model for calling short indels in tandem repeats (2.6.4).

2.6 Detecting indels in tandem repeat loci using short read sequence data

To investigate whether a tandem repeat is different in length in a sequenced sample compared to the reference genome, we compared the distribution of MPERS for read pairs that map on either side of a given repeat locus to the calculated distribution of MPERS for all read pairs in an individual's genome that maps uniquely across a given repeat length (see figure 2.2). Put simply, a shift in the MPERS distribution at a given locus to the right suggest that the repeat locus is smaller in the sample than in the reference genome, while a shift to the left suggests it is longer. Based on this understanding of how paired end mappings work across indels, our method iterates through all plausible allele configurations for a diploid genome at each short tandem repeat locus and estimates the most likely lengths of the two copies in a sequenced sample by using a maximum Bayesian posterior approach.

2.6.1 Background on indel detection using paired end sequence data

Before delving into the intricacies of determining the repeat length based on the the distribution of MPERS, assume first that the length of the sequence fragments in a library could be held constant at some chosen value. If a sequenced tandem repeat locus was the same length in a sample as in the reference genome, the MPERS for a read pair sequenced from either end of a fragment containing that locus should be equal to the chosen fragment length for that library. However, when sequence is removed from a repeat locus in a sample relative to the reference genome – as is the case for deletions – the MPERS for a read pair sequenced from either end of a fragment containing the locus will be longer than the chosen fragment length for the library. This happens because when sequence is removed in a sample, the reads of a spanning read pair are mapped further apart than expected. The actual fragments coming from the fragment library have not changed in length, only the sequence between the reads has changed relative to the reference. The same principle holds true in the opposite direction

for insertions: the reads of a spanning read pair are mapped closer together, and so the MPERS is smaller than expected. Figures 2.2 and 2.3 illustrate how such a shift would occur by comparing two scenarios where the sequenced sample has either the reference repeat length allele or a deletion.

In reality, the fragments in a given library have a distribution of lengths approximately centered at the chosen fragment length for the library. Thus, to identify an indel in a repeat locus, we must test whether the distribution of MPERS for spanning read pairs spanning the locus matches better with a different distribution of MPERS than that of the distribution of sequence lengths in the fragment library (see section 2.6.3.1). Ideally, shifts in the mean MPERS across a sequenced repeat locus to the left and right compared to the mean MPERS for a fragment library are indicative of an insertion or deletion, respectively.

2.6.2 The empirical distribution of MPERS for read pairs that span a STR locus of a given length, $P_l(\mathbf{M})$)

The main underpinning of our model for detecting indels in STRs involves examining the distribution of MPERS for read pairs whose two reads map on either side of a repeat locus (spanning read pairs). When looking at STR loci, spanning read pairs are independently mapped around an STR but are constrained by the fact that they must be sequenced from a fragment that is at least as long as the STR with enough bases outside the repeat to map uniquely to the flanking sequence. This inevitably has the effect that the longer the STR locus is in the sample which was sequenced, the higher the mean MPERS of its spanning read pairs will be (as illustrated in figure 2.4). As well as an increase in the mean MPERS for longer STRs, the number of spanning reads at a given locus is reduced as the STR increases in length. More directly, as the repeat tract approaches the length of the chosen fragment size, the proportion of reads capable of spanning the repeat locus diminishes in line with the size of the repeat length which is independent of the sequence coverage (2.5a). This has the reciprocal effect of increasing the number of hanging/anchoring reads around an STR as the repeat length increases. This trade off from spanning mate pairs to hanging/anchoring

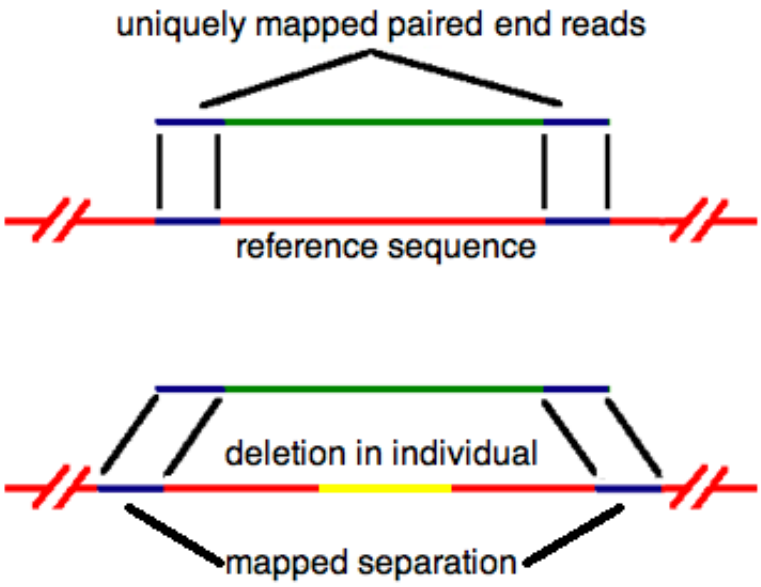


Figure 2.2: Graphic of mapped paired end read alignments of an individual whose locus matches the reference (top) and whose locus contains a deletion in respect to the reference (bottom). The blue paired end reads at top align at the exact distance one would expect to observe given the fragment length of a sequenced library, which is indicative of the individual having the same locus length as in the reference. The blue paired end reads at bottom, however, map further apart due to a removal of bases in the sequenced individual (yellow line). The removal of bases in the sequenced individual will therefore cause all mapped paired end reads across this locus to appear to map further apart than the expected MPERS for the given library.

reads is more or less linear with increasing repeat length until the repeat tract surpasses the fragment length library size where there are no longer any spanning reads and the number of hanging/anchoring reads remains constant (2.5b). This restriction represents the main limiting factor in the robustness of our approach to genotyping STRs in a deep sequenced individual. Unlike many problems in sequence assembly and resequencing analysis where additional sequencing helps, the problem of assaying longer STRs can only be rectified by creating a new library with a longer fragment size. Knowing the distribution of MPERS across varying repeat lengths was crucial to the efficacy of our model. Because the dis-

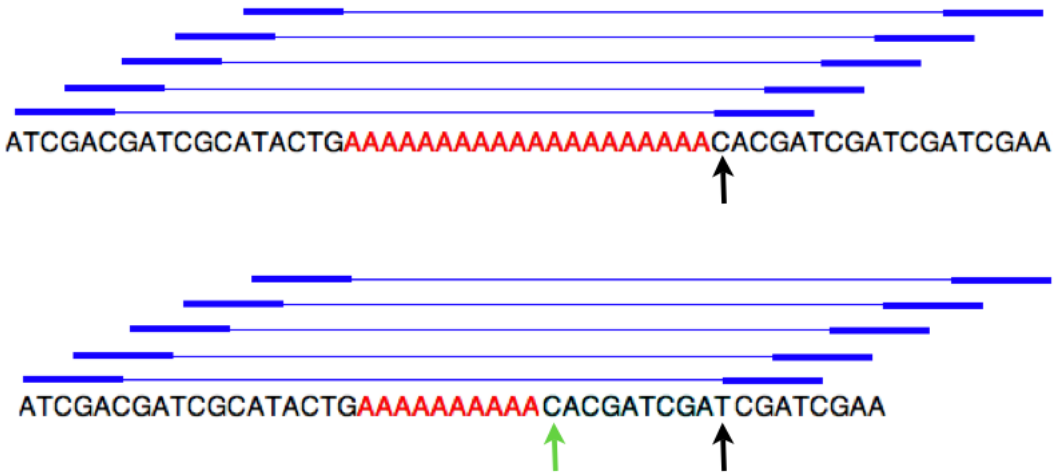
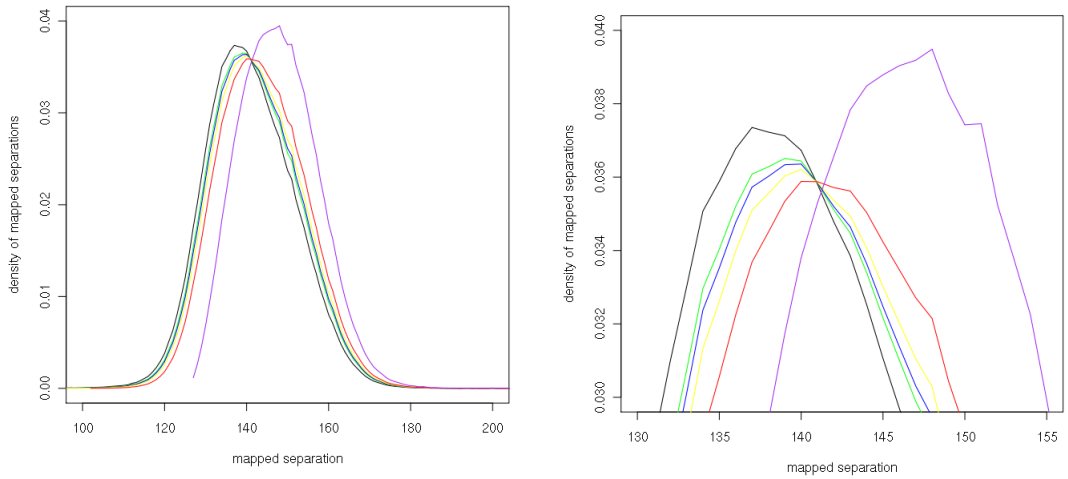


Figure 2.3: Mapped paired end reads sequenced from an individual whose reads align to both the reference repeat length (top) as well as a deletion in the repeat tract in respect to the reference (bottom). The bottom most read pair in the top illustration (blue, closest to the sequence) has a span of 43 bp that encapsulates a poly-A chain of length 20 bp. This mate pair's right read begins one bp to the right of the repeat tract (black arrow, base C). As reads from this library are only 5 bps in length, it is not possible to directly sequence across this repeat tract and determine the overall length, but as the read maps at the distance one would expect given the fragment length library, we can assume that the sequenced individual's repeat length is the same as that of the reference length. The sequence at bottom contains a deletion of 10 bp in the poly-A repeat tract. This deletion effectively causes the bottom most read to map 10 bps downstream of the repeat (from the green arrow to the black arrow) making the MPERS appear larger than they actually are when compared to other MPERS in the same library. This anomalous mapping would be indicative of there being a 10 bp deletion in the repetitive tract.

tribution of MPERS naturally drifts upwards as the repeat length increases, it was paramount we know what the true distributions of MPERS across varying repeat lengths were, otherwise we would make numerous false positives in the form of deletions.

Our initial approach in determining the distribution of MPERS across varying repeat sizes was to amalgamate all repeats within the genome of a given size into groups and the distribution of reads across these groups were calculated. As our

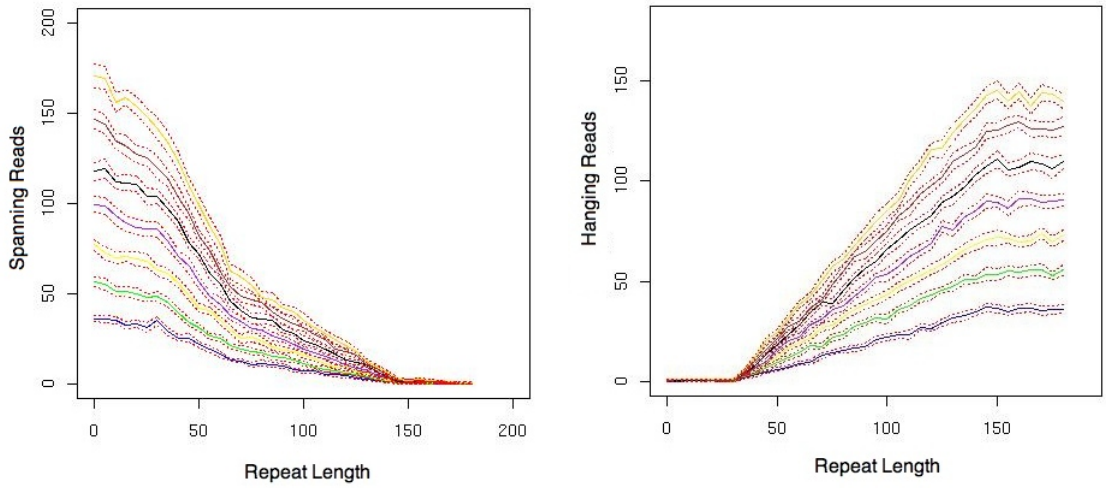


(a) Distribution of MPERS for library g1k-sc-NA12878-CEU-1 across differing repeat lengths

(b) Inset of MPERS peaks for graph (a)

Figure 2.4: Empirical distributions of MPERS for individual NA12878 library g1k-sc-NA12878-CEU-1. (a) Distribution of MPERS for all read pairs used in calculating the empirical distribution for library g1k-sc-NA12878- CEU-1 (black) as well as the subset distributions of MPERS for read pairs that would span a specific repeat locus of length 25 (green), 50 (blue), 75 (yellow) 100 (red) and 125 bp (purple). (b) Close-up of the distributions peaks illustrating the right tending of the MPERS distribution as the repeat locus length increases.

model needs the values of all possible repeat lengths, this posed a problem as many of the longer repeat lengths were not extremely prevalent in the genome. This method also had the problem that the mapping of reads across repeats in the genome were not always uniform and as expected. If there were proximal repetitive regions to a given repeat of a known length, they could cause the reads to map further than expected due to the inability of shorter paired end reads to map uniquely across both the tandem repeat as well as the adjacent repetitive sequence which in turn would throw off our calculation of the empirical distributions. Lastly, the regions in the genome might not match the reference length in the sequenced sample. For example, if a site in the reference measured 60 bp and the sample sequenced had a deletion of 21 bp, the read pairs from that



(a) Graph of spanning read pairs

(b) Graph of hanging reads

Figure 2.5: Simulation results of the number of spanning (a) and hanging (b) reads across different coverages and repeat lengths from a constant fragment length library. Graphs (a) and (b) represent the number of spanning read pairs and hanging reads, respectively, observed when simulating a repeat tract of 0 to 200 bp by increments of 5 bp (y-axis) at different coverages (10 to 40x by increments of 5, bottom to top) of a fragment length library of 150 bp and a standard deviation of 0 and then mapping simulated paired end reads from the sequence and mapping them back to the sequence from which they were just sequenced from. Thirty simulations were conducted for each repeat length, coverage with the dotted lines representing 1 standard deviation above and below the mean number of spanning/hanging reads observed for a given repeat length, coverage pair. As the repeat tract approaches the fragment length library size, the number of spanning reads approaches zero and no spanning reads are observed after this point. Hanging reads work in exactly the opposite direction where their numbers increase up to the fragment library length, but then level out once the repeat tract increases above the fragment library size. For full discussion on how the simulations were performed, see 2.8

sample's locus would be used in calculating the distribution of MPERS for repeat lengths of 60 bp, not 39 bp. These concerns led us away from calculating the distributions of MPERS for repeat lengths directly from spanning reads in the genome to a more theoretically-based approach which used the distribution of MPERS for the uniquely mapped reads we had already gathered (see section 2.4).

The expected distribution of MPERS for read pairs that span a STR of a given length, $P_l(M)$, was calculated by looking through the entire set of read pairs in the genome wide screen and generating a subset of read pairs whose MPERS were of sufficient length to map uniquely on either side of a repeat locus of a given length, l. Only read pairs whose MPERS were two bp longer than a repeat locus's length were added to the subset. This criterion assured that the MPERS were of sufficient length that one bp of each read could map outside the repeat locus, thus anchoring it in the adjacent unique sequence.

We iterated through every possible l that could be spanned by a fragment library (10 bp to 6 standard deviations above the mean MPERS of the fragment library) and generated an empirical distribution; we did not generate any distributions for $l < 10$ bp as they were considered of insufficient length.

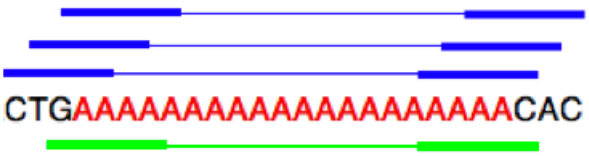


Figure 2.6: Cartoon representation of actual mapping positions of two paired end reads across a poly-A repeat of length 20 bp. The blue paired end read of MPERS 24 bp has three unique positions it can map to and still uniquely span the repeat, while the green paired end read of MPERS of 22 bp has only one.

To calculate the number of possible mapped positions that a read pair could have given that it has a MPERS value of $M = m$ and spans a repeat locus of length l, we use the following equation

$$n_l(m) = m - l + 1 \tag{2.1}$$

For example (as shown in Fig 2.6), if a read pair has a MPERS of $m = 24$ bp and is spanning a 20 bp repeat locus, there are three possible mapped positions that the read pair could have, $m_{20}(24) = 3$. The number of unique mappable positions is important to consider because much longer MPERS will have a lower probability of being observed in the genome, but will actually have a much higher chance of spanning a repeat. Now knowing the $n_l(m)$, it is simply a matter of exhaustively looking at all possible mappable positions for each MPERS in a subset of read pairs across a repeat length and determining the probability of observing a spanning read pair of a given MPERS. This probability was calculated by multiplying the $n_l(m)$ by the frequency of observing a spanning read pair of length m ($F(m)$) from a given library. The $P_l(M)$ for a library was calculated as follows:

$$P_l(M = m) = \frac{n_l(m) \cdot F(m)}{\sum_{m'} n_l m' \cdot F(m')} \tag{2.2}$$

where the denominator in equation 2.2 normalizes the estimated probabilities.

2.6.3 Estimating the genotype of a tandem repeat locus

When estimating the size of a putative indel, as discussed earlier, we consider that for longer repeat loci there is a higher probability of observing spanning read pairs with higher MPERS values; that the true length of the repeat locus in the individual will affect the distribution of MPERS observed when read pairs sequenced from fragments that contain that locus are mapped to the reference genome. The true distribution of MPERS for a given locus plays an important role in ascertaining the correct allele length when maximizing the posterior probability of a locus containing an indel of a given size based on the observed paired end reads spanning a locus (see 2.6.3.1).

2.6.3.1 Rationale behind analysing MPERS distributions to detect indels in STR loci

If a sequenced STR locus is the same length in a sample as in the reference genome, the distributions of MPERS for paired end reads sequenced from either

end of a fragment containing the locus should be as given by equation 2.2. However, if there is a deletion in this individual's repeat locus relative to the reference genome, the MPERS for a read pair sequenced from fragments spanning the locus will tend to be greater than expected on the order of the size of the indel. For example, consider an individual with a homozygous insertion of 15 bp in a repeat locus relative to the reference genome (indel size $i = 15$). The mean MPERS for the read pairs that span the repeat locus will be shifted approximately 15 bp to the left (or 15 bp shorter than expected). If the size of the true indel length is then added to the MPERS for each of the spanning read pairs, the resulting distribution would align with the true underlying distribution given by equation 2.2, with l increased by i

$$P_{l+i}(m+i) \tag{2.3}$$

Using the same example as before, assume the repeat locus length in the reference genome was 60 bp; therefore the length of the repeat locus in the sample's copies would be 75 bp. Because of this, we must compare the distribution of MPERS for paired end reads spanning the locus to the probability distribution of MPERS for all spanning paired end reads that span repeat loci that are 75 bp in the sample sequenced.

The inherent problem with equation 2.3 was that it only considered a single allele (haploid), precluding the model's ability to make correct genotype calls for individuals that were heterozygous at a locus. If the individual is homozygous at a STR locus, then all read pairs that span the locus will be drawn from two identical distributions, whereas at a heterozygous locus, there will be two distributions that a spanning paired end read can come from with a 50% probability that a paired end was drawn from each of the two distributions corresponding to the separate copy lengths. We note that the actual probability of a paired end read being drawn from an allele is contingent upon the repeat length in the sequenced sample, and when different – as is the case for heterozygotes – the smaller of the two alleles has a marginal gain in the probability that a read was drawn from it (independent of its MPERS) as the number of sites a paired end

read can uniquely map to increases (as stated in equation 2.1). But, as this gain was negligible for most cases as the difference in repeat lengths in each of the copies was rarely observed to be extremely different (see section 2.6.4), it was ignored and the probabilities of drawing from one allele or the other were set equal.

Ultimately, it was necessary to be able to genotype any of the following scenarios: a locus which is homozygous with two reference alleles ($i_1 = 0, i_1 = i_2$), homozygous with two non-reference alleles ($i_1 \neq 0, i_1 = i_2$), heterozygous where one allele matches the reference length ($i_{\{1,2\}} = 0, i_1 \neq i_2$) or heterozygous where neither alleles matches the reference length ($i_1 \neq i_2 \neq 0$).

As the model iterates through all possible indel size genotypes (i_1 and i_2), for computational ease it was important to constrain our predictions to a sensible range. Having initially assayed tandem repeats using capillary reads (see section 2.5), we knew that a majority of all indels for a given motif length fell within ± 10 repeat units of the reference length and were of multiples of the motif length for shorter repeat motif lengths; of the 155,676 calls made from capillary reads in the Trace Archive for repeat motifs of length thee (triplets), only 56 (0.03%) fell outside the range of [-30,30] and 2069 (1.3%) were not multiples of three. From here, we could now calculate the probability that the observed MPERS came from distributions and reads which corresponded with the underlying true repeat lengths in the sequenced sample's two distributions corresponding with the putative genotype call $\{i_1, i_2\}$ which relate them to the underlying repeat length of the two copies ($P_{l+i_1}(M = m + i_1)$ and $P_{l+i_2}(M = m + i_2)$). The likelihood of the data given the hypothesized genotype can then be calculated as

$$L_{l+i_1, l+i_2} = \prod_{s \in r} [\frac{1}{2} P_{l+i_1}(m_s + i_1) + \frac{1}{2} P_{l+i_2}(m_s + i_2)]$$

where r is the set of read pairs, s, spanning the locus, and m_s is the MPERS of read pair s. We then maximized this likelihood and arrive, hopefully, at the true

genotype

$$\arg\max_{i_1,i_2} L_{l+i_1,l+i_2}(i_1, i_2|r) = \{\hat{i_1}, \hat{i_2}\}$$

In practice, however, the maximum likelihood estimate was usually incorrect due to the natural variation in the distribution of MPERS. Without a prior, the model ran the risk of overcalling false positives. This problem was directly observed during our simulations to check the proof of concept for our model (see section 2.8.1). The heatmap in figure 2.7 illustrates this concept of over fitting the data which will occur without the necessary priors in place.

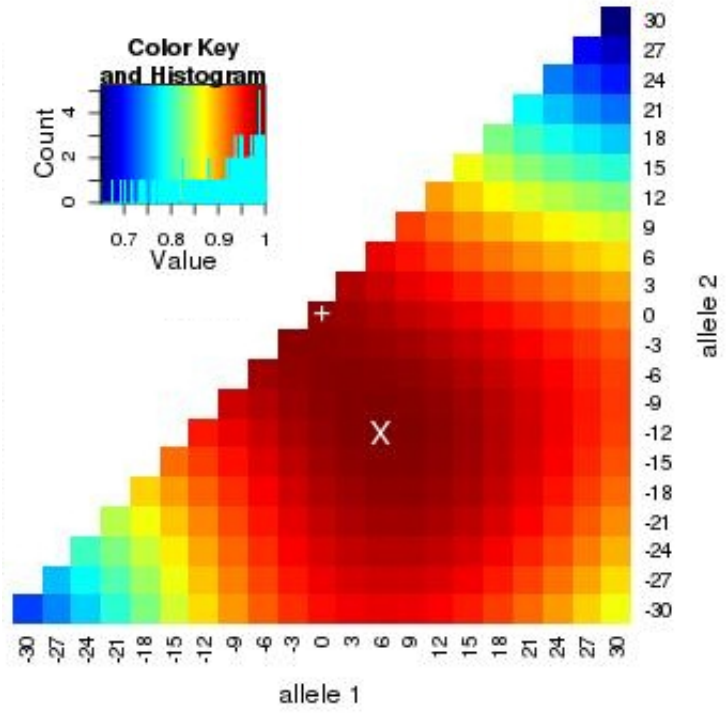


Figure 2.7: Heatmap of likelihoods at a selected repeat locus of length 60 bp from a simulated homozygous reference genotype with average base pair coverage 15x. In the simulation, the maximum likelihood estimate determined the genotype to be {6,-12} (white X), where the actual genotype was reference (white +). The distribution of likelihoods is unimodal, centering around the incorrect genotype call X caused by the random variation in the MPERS of a fragment library. Simulation methods are discussed in section 2.8.

From this understanding that the maximum likelihood would not suffice in correctly genotyping a STR, we estimated the probability of genotype $\{i_1, i_2\}$ at a given locus by employing a Bayesian approach which incorporated a genotype prior that will be discussed below.

Bayes' theorem states that the probability of A given B is equal to the likelihood of B given A times the prior probability of A divided by the probability of B.

$$P(A|B) = \frac{L(B|A)P(A)}{P(B)} \tag{2.4}$$

To find a sample's genotype, we calculated the proportional posterior probability of an indel pair $\{i_1, i_2\}$ at a given locus by multiplying the likelihood of the set of paired end reads (r) by the prior probability of the putative genotype

$$P_{l+i_1,l+i_2}(i_1, i_2|r) \propto P(i_1, i_2|k) \cdot L_{i_1,i_2} \tag{2.5}$$

$$L_{i_1,i_2} = \prod_{s \in r}[\frac{1}{2}P_{l+i_1}(m_s + i_1) + \frac{1}{2}P_{l+i_2}(m_s + i_2)]$$

where $P(i_1, i_2|k)$ is the prior probability of genotype $\{i_1, i_2\}$ given its motif repeat length is k. The methods for which we estimate the prior probabilities are described in section 2.6.4. As we were interested in ascertaining the most probable genotype, we searched for which indel pair maximized the proportional posterior probability of equation 2.5.

$$\arg\max_{i_1,i_2} P_{l+i_1,l+i_2}(i_1, i_2|r) = \{\hat{i_1}, \hat{i_2}\}$$

This calculation was performed in log space to rectify the problem of numerical underflow in determining the genotype which maximized the posterior probability.

Because many deeply sequenced individuals are sequenced from multiple libraries, it is important to combine the shared information across libraries in determining the correct genotype. As the signal for the underlying true repeat length is interpreted the same by any spanning read pair sequenced from a library, we were able to combine the information from different libraries by assuming the sequencing of all libraries (as the same as paired end reads in a library) are independent of one another, and then by taking the product of equation 2.5 for each library, we were left with

$$P_{l+i_1,l+i_2}(i_1, i_2|r) \propto P(i_1, i_2|k) \cdot \prod_{b \in t} L_{i_1,i_2,b}$$

$$L_{i_1,i_2,b} = \prod_{s \in r_b}[\frac{1}{2}P_{l+i_1,b}(m_s + i_1) + \frac{1}{2}P_{l+i_2,b}(m_s + i_2)]$$

where t is the list of libraries sequenced from an individual, r_b is the set of spanning read pairs for library b and $P_{l+i_{\{1,2\}},b}$ are the MPERS distribution for library b.

2.6.4 Prior Probabilities

In estimating the genotype priors for our model, it was necessary to first ascertain the distribution of indel sizes across the genome before estimating the probabilities of genotype configurations. The priors were calculated using the haploid calls made from the capillary data in NA18507 as described in section 2.5. Due to the low coverage of capillary reads for NA18507, we were usually able to infer only one copy length per STR locus. Out of the 17,181 triplet repeat loci in the autosomes at which we inferred at least one copy length, only 206 sites had evidence for two separate repeat lengths. This does not mean that the probability of observing a heterozygous locus is 1.2%, but that there was not sufficient sequencing to know the true genotype at each locus. Because of this, we used each call as a haploid to estimate the prior probability of observing an indel of size i bp in a STR locus relative to the reference genome. The priors were conditioned upon which family they belong to in regards to their repeat length motif unit size of k bp. The distribution was estimated from the total number of alleles observed in NA18507 that contain indels of size i bp divided by the total number of alleles observed in NA18507 containing indels of any size (including no indel, $i = 0$). The value $P(i|k)$ was therefore calculated as

$$P(i|k) = \frac{F(i|k)}{\sum_{\forall i'} F(i'|k)}$$

where $F(i|k)$ is the number of single allele calls observed in individual NA18507 that contain indels of size i bp for a repeat length motif of size k.

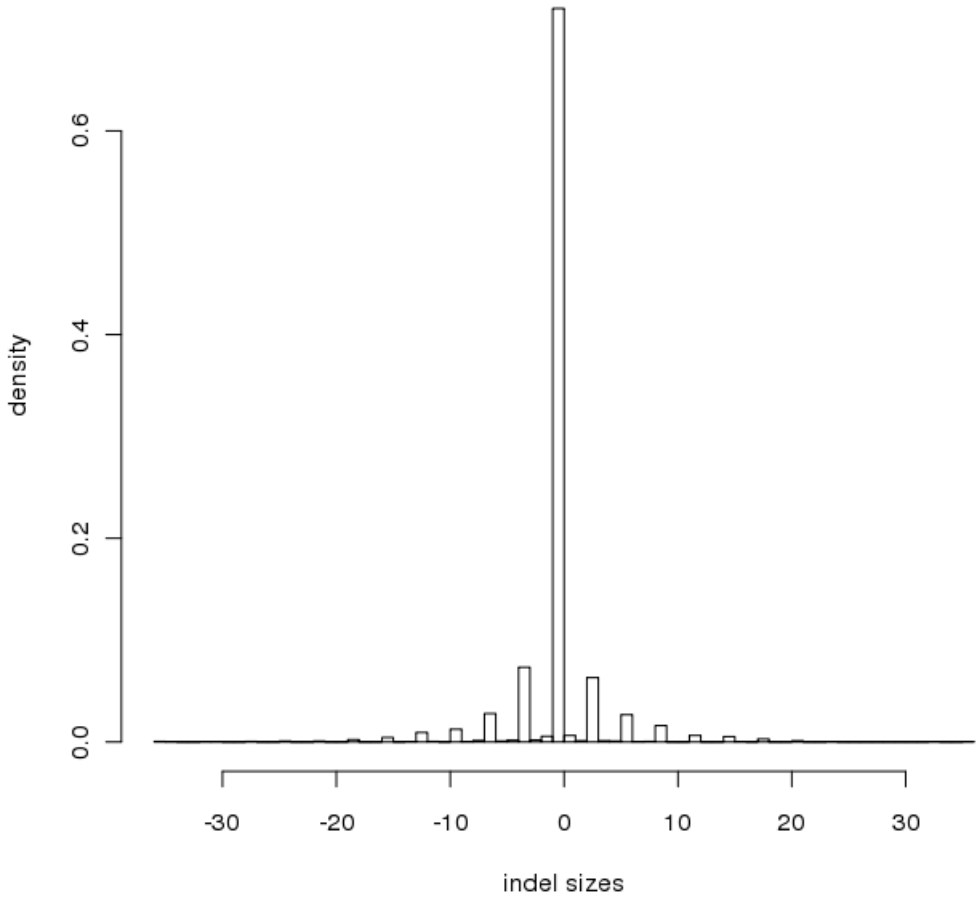


Figure 2.8: Prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads. This distribution is based on 3,435 calls; 2,474 reference calls (72.0%) and 961 indel calls (28.0%).

For our prior, we choose a call set whose values did not put as much weight on the prior as to force all true calls to be reference calls. In total, this call set was comprised of 3,453 autosomal calls (3,225 single allele calls, 105 heterozygote calls). However, a problem came to light when we looked at the distribution of prior probabilities for our indel distribution data set. When the number of insertions versus the number of deletions were compared for the same absolute size indels, there is a bias towards observing deletions over insertions.

Because the reference length is selected at random, we believe this bias is not biological, but in fact an artifact of the calls made by the capillary alignment.

Rectifying this problem was completed simply by averaging between same magnitude insertion and deletion calls. For example, if $P(i = -6|k = 3) = .3$ and $P(i = 6|k = 3) = .2$, the estimated prior probability for a structural variant of magnitude 6 was generally calculated as 0.25 using

$$P(|i||k) = \frac{P(-i|k) + P(i|k)}{2}$$

which yielded a symmetric distribution of prior probabilities mirrored across the reference allele length (figure 2.9). Indels that were not of magnitudes in multiples of the repeat motif length were proportionally pooled into their nearest two adjacent bins to remove any intermediary calls. Because the mutation rate at STR loci varies between sites and can be quite high, there is a significant probability of multiple alleles at a locus, and we cannot derive the distribution of genotypes from the distribution of indel sizes by assuming Hardy-Weinberg independence. We also were not able to estimate the genotype distribution from the NA18507 capillary alignment data, because the depth was inadequate to reliably sample both alleles (as we only observed 206 heterozygous sites in the large call data set). Therefore we based our genotype prior heuristically on the following assumptions:

1. The most likely genotype is a homozygous genotype where both copies of the repeat locus in the sample are the same length as the repeat length observed in the reference genome, $\{i_1 = 0, i_1 = i_2\}$.

2. The second most likely genotype is heterozygous with one reference allele length and one non-reference (indel) allele length, $\{i_{\{1,2\}} = 0, i_1 \neq i_2\}$.

3. The third most likely genotype is a homozygous indel in respect to the reference genome length, $\{i_1 = i_2, i_1 \neq 0\}$.

4. The least likely genotype is a heterozygous genotype where both alleles differ in length from the reference length, $\{i_1 \neq i_2 \neq 0\}$.

Based on these assumptions, we estimated the relative prior probabilities of the non-homozygous reference genotypes as follows (scaled to a value of 1 for the homozygous reference genotype): for a heterozygous genotype with one reference

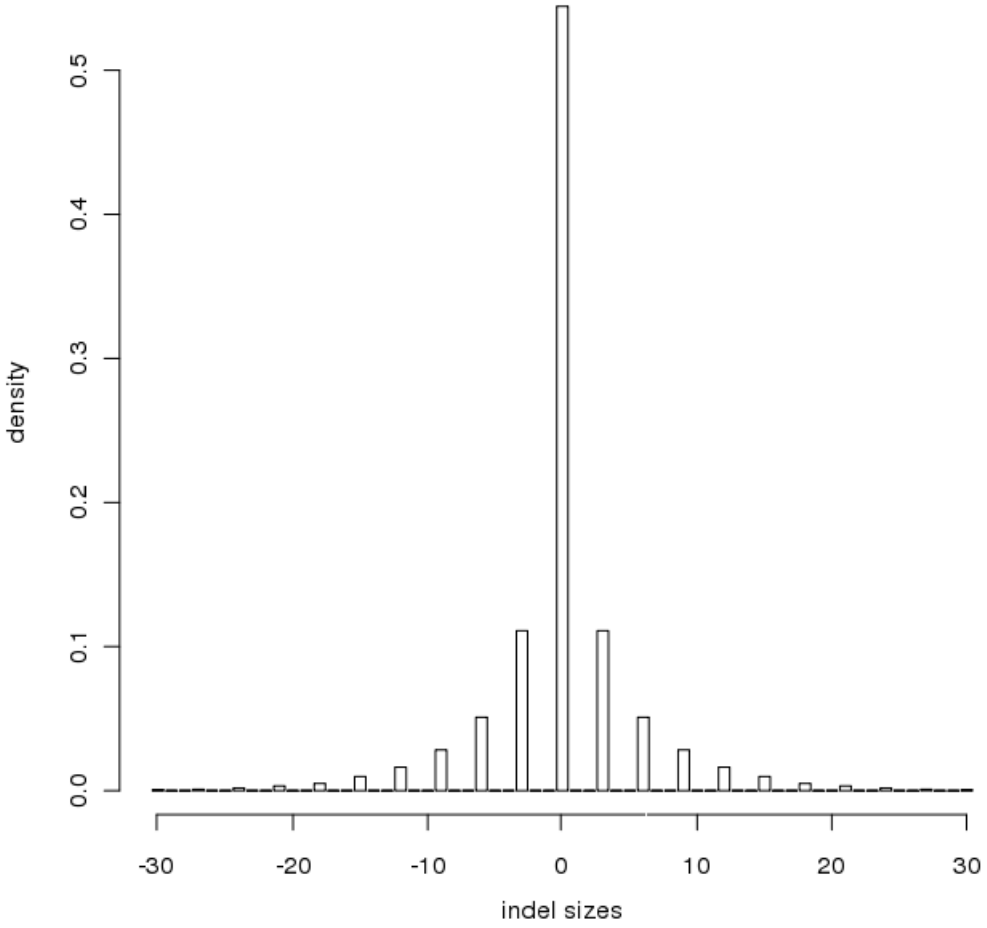


Figure 2.9: Symmetric prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads.

allele $\{i_{\{1,2\}} = 0, i_1 \neq i_2\}$ we used the value $P(i_{\{1,2\}}|k)$ (the probability of observing an indel of size i), for a homozygous indel $\{i_1 = i_2, i_1 \neq 0\}$ we used $0.5 \cdot P(i_1|k)$, and for a heterozygous genotype with two non-reference alleles $\{i_1 \neq i_2 \neq 0\}$ we used $P(i_1|k) \cdot P(i_2|k)^{0.5}$, where the absolute value of i_1 is larger than that of i_2. This prior assured that the calls would be more accurate than simply assuming the two copies repeat lengths were independent of one another. When graphed, the prior probability space illustrates the areas we would expect to see more calls when assaying a number of repeats across a genome. Figure 2.10 is a representation of the prior probability space of repeat length motif $k = 3$.

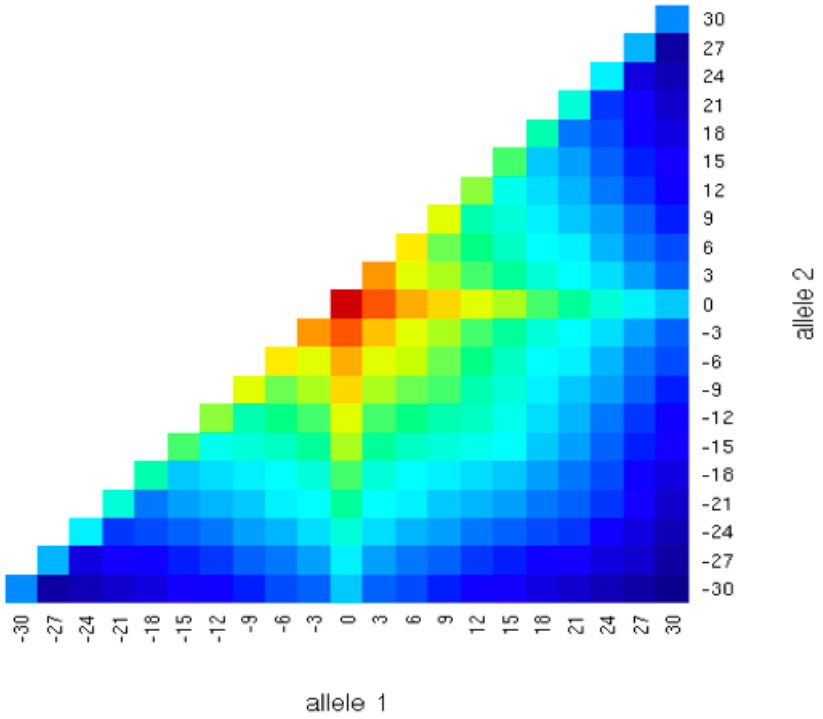


Figure 2.10: Heat map of the estimated prior probabilities for the varying genotypes of a triplet repeat locus in an individual (shown in log space). The confirmation of the probability space illustrates the assumptions made about which genotypes will be more likely than others. The most probable genotype is the homozygous reference (0,0; red), followed by a heterozygote with one reference allele (horizontal and vertical lines where allele 1 or allele 2 equals 0), a homozygous indel (top diagonal line) and lastly a heterozygote with neither allele matching the reference length.

Literature on the mutability of tandem repeats generally agrees that the composition of the repeat motif (v), as well as the repeat locus length (l) in the reference can either increase or decrease the repeat locus's likelihood of undergoing an insertion or deletion event (Ellegren [2004]). Ideally, the prior probability of a locus would be conditioned on both the v and l ($P(i_1, i_2|k, v, l)$) but there was insufficient data to incorporate this information into our prior.

2.6.5 Odds ratio and normalized posterior

As a measure of our confidence in a genotype call for a repeat locus in a sample, we calculated the ratio of the posterior probabilities of the maximum posterior call to the reference homozygous call. This ratio gave us an idea of which calls had the most evidence that a locus was non-reference. In a large call data set – as is the case for the human genome – the odds ratio gave us a good indication of which loci had more evidence for our call to be correct compared to all other calls which may serve as a filter after determining which of our calls are correct through validation.

$$\text{odds ratio} = \frac{P_{l+\hat{i}_1,l+\hat{i}_2}(\hat{i}_1,\hat{i}_2|r_b)}{P_{l,l}(i_1=0,i_2=0|r_b)} \tag{2.6}$$

For later analysis, we needed to calculate the full posterior probability of our calls as opposed to the proportional posterior which sufficed in determining the indel pair which maximized equation 2.5. This value was calculated straightforwardly as

$$P_{l+i_1,l+i_2}(i_1,i_2|r_b) = \frac{P(i_1,i_2|k)\cdot L_{l+i_1,l+i_2}}{\sum_{i'_1,i'_2} P(i'_1,i'_2|k)\cdot L_{l+i'_1,l+i'_2}}$$

$$L_{l+i_1,l+i_2} = \prod_{r\in r_b} \frac{1}{2}P_{l+i_1}(M=m_s+i_1) + \frac{1}{2}P_{l+i_2}(M=m_s+i_2)$$

$$L_{l+i'_1,l+i'_2} = \prod_{r\in r_b} \frac{1}{2}P_{l+i'_1}(M=m_s+i'_1) + \frac{1}{2}P_{l+i'_2}(M=m_s+i'_2)$$

where the denominator normalizes the probability which we previously omitted in equation 2.5. Due to our omission of the denominator, we calculated the proportional probability in log space to avoid numerical underflow. Incorporating the denominator added the complexity of what to do with the log of a summation – a non-trivial task. Luckily, we were able to locate a solution to this problem known generally as the 'logsumexp trick.' The logsumexp trick is easily found through

any google search, as well as in many statistical analysis books (Durbin [1998]) to answer the problem of underflowing a computer's resources when calculating the normalization constant in Bayes' theorem. A relatively straight forward solution, the logsumexp trick exploits the inherent logarithmic property that raising a log to its base yields simply the value of the number.

$$x = e^{\ln(x)}$$

From this, the log sum can be calculated directly as follows

$$\text{logsumexpexp}(a_t) = \log \sum_t \exp^{a_t}$$

$$\log \sum_t \exp^{a_t} = \log \sum_t \exp^{a_t} \exp^{A-A} = A + \log \sum_t \exp^{a_t - A}$$

where $A = max\{a_t\}$. Now able to calculate the posterior probability for each genotype call, it became a matter of simply setting up the ratio between the genotype call which maximized the posterior probability to that of the reference genotype call (see equation 2.6).

2.7 Software

The software to implement this model, called STRYPE, is available as an end user package for genotyping tandem repeats. Source code and supplementary material (including a test data set for individual NA18507) can be found at
https://sourceforge.net/projects/strypecode/
Individual NA18507 was chosen as a test set to minimize the number of files that needed to be downloaded by the user to test the program. As each library needed its own series of distributions specific to its fragment size library, NA18507 was a perfect sample as it was sequenced to a deep coverage by a single library.

2.8 Simulations

Before running the model on any real data, it was important to test the proof of concept before engaging in any further analysis. This test was conducted using the alignment tool MAQ which comes with an added feature that allows users to generate a simulated set of paired end reads. These reads are drawn from a Gaussian distribution whose mean and standard deviation are input by the user. As well as the library's fragment size parameters, MAQ's input includes the length (b) of each of the paired end reads (chosen as 35 bp for this simulation) as well as the number of paired end reads (z) to be simulated. This input was ancillary to the more quoted statistic of bp coverage (c); the number of times a base is sequenced by the reads in a sample. Determining the c of a sample is completed simply by multiplying the number of reads by their read length and dividing by the sample length (in bp)

$$c = \frac{2 \cdot b \cdot z}{g} \tag{2.7}$$

where the coefficient of two is for the fact that the number of reads simulated are in pairs and must be considered separate when calculating c.

Next, we selected a region of 1,800 bp from the genome of *Streptococcus suis* that contained no repeat tracts. This sequence (in fasta format) was then split in half at position 900 at which we introduced a STR of a predetermined length. We chose the STR to be of motif CAG and of pure tract with a length that was a multiple of the motif size ($k = 3$). Each genotype scenario (described below) was simulated and our model's accuracy was scrutinized.

The genotype determined which simulated sequence the reads were generated from and the reference repeat length determined which simulated sequence they were mapped back to. The MAQ simulations were run for a given μ and σ as well as c which incorporated both the 1,800 bp reference sequence plus the additional repeat length in the sequence. For simplicity, all simulation examples described

below will have the following parameters: l is 60 bp, c is 40x, μ is 200 bp and std is 10% of the mean (20 bp).

2.8.1 Reference

As the most observed genotype when looking across all STRs in a sequenced genome (see section 2.9), it was important that our simulations prove that the prior distribution would alleviate the problem of making false positive calls based on the natural variation in the fragment length library. As described in section 2.6.3.1, a simple maximum likelihood would cause there to be numerous false positives and downgrade the efficacy of our model. However, the addition of a prior based both on the magnitude of the indel calls as well as their genotype should bring our call accuracy more in line with the truth.

Simulating reference genotypes were the most straight forward process as they did not rely on generating sequence for multiple samples. Using the above prescribed user input, 60 bp of CAG sequence was inserted into the truncated region of *S. suis* starting at position 900. This fasta sequence was then input into MAQ simulate and reads were simulated corresponding to the user's input. The simulated paired end reads were then mapped back to the sequence and the map file alignments were then run through our model. We noted that more times than not, the maximum likelihood estimate would place the genotype off the reference (782 of 1000 simulations), but the addition of the prior decreased this number to 9 – a 0.9% false positive rate. The set of spanning paired end reads is consistent with the underlying MPERS distribution from which they were sampled (figure 2.11).

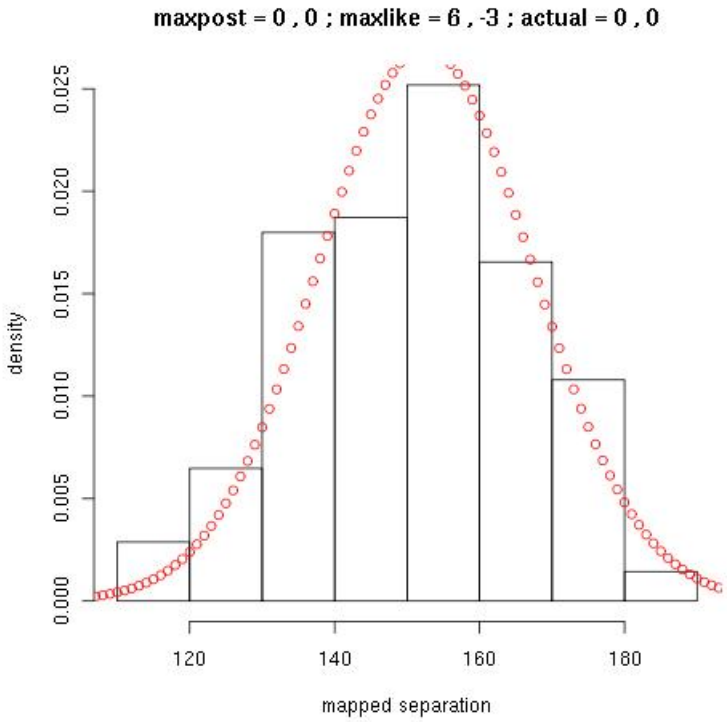


Figure 2.11: Histogram of spanning paired end read separations and MPERS distribution for a reference genotype simulation. The histogram of spanning paired end read separations across a simulated repeat tract coincide distinctly with the distribution of MPERS you would expect to observe given the sample is homozygous at a locus with repeat length l.

2.8.2 Homozygous indel

The second least complex simulation scenario, the homozygous indel, required only a single additional step. Unlike the reference genotype, however, two sequences were generated to emulate the scenario of a homozygous indel at a STR locus. To start, a reference sequence was generated to which the MAQ simulated reads would be mapped. An additional sequence was generated which corresponded with the length of the true repeat tract

$$l_{new} = l_{reference} + i \tag{2.8}$$

For instance, a homozygous deletion of -21 bp in a STR of length 60 bp would mean the paired end reads would be simulated from a sequence which contained a repeat tract of 39 bp. The reads simulated from the shortened repeat sequence sample would then be mapped back to the reference containing a repeat length of 60 bp. When graphed, this would look as if the set of spanning paired end reads were mapped 21 bases further apart than what would be expected given the reference repeat locus length (figure 2.12). In the example shown, the maximum likelihood genotype was {-18,-21}, but the maximum posterior probability genotype was {-21,-21} which is correct. The power of our model to detect indels is contingent upon the underlying genotype; as the genotype diverges more from the reference, the more power our model has for correctly genotyping the individual.

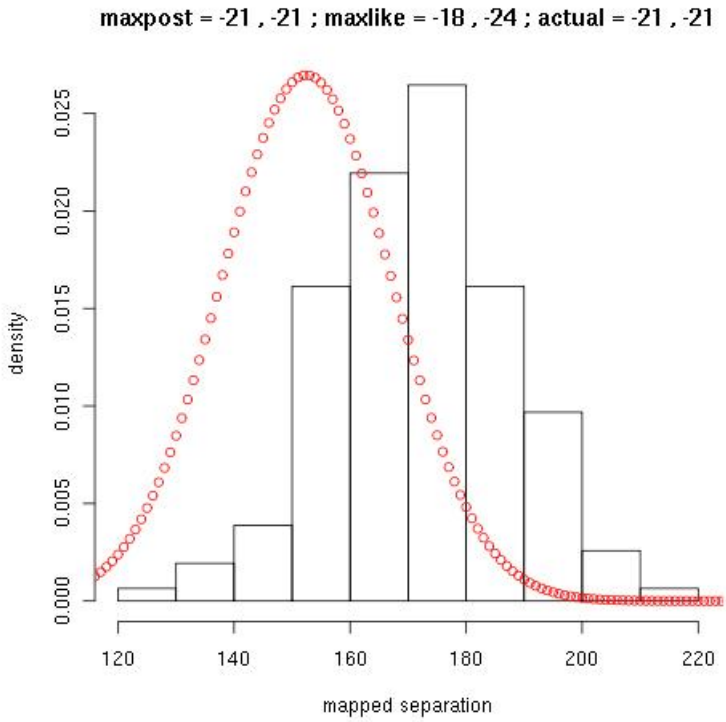


Figure 2.12: Histogram of spanning paired end read separations across an individual whose repeat length is 21 bp shorter than that in the reference graphed against the MPERS distribution for a reference length genotype. The mean MPERS for the spanning paired end reads is therefore shifted approximately 21 bp to the right.

To assess the accuracy of our model in calling homozygous indels, we've simulated each of the plausible homozygous indels within a biologically relevant range ([-30,30] by units of three bp) 50 times and checked our model's accuracy. The values for these simulations are listed in table 2.2.

Simulation accuracy statistics for homozygous indels			
Indel size	Genotype		Number of genotype calls
30	30	30	24
30	27	27	15
30	24	24	11
27	24	24	26
27	27	27	12
27	30	30	6
27	21	21	6
24	24	24	22
24	21	21	16
24	18	18	7
24	27	27	5
21	18	18	23
21	21	21	20
21	15	15	5
21	24	24	2
18	15	15	22
18	18	18	15
18	12	12	6
18	21	21	4
18	30	0	1
18	24	24	1
18	24	0	1
15	12	12	20
15	15	15	19
15	18	18	6
15	9	9	3
15	24	0	1
15	21	0	1
12	12	12	23
12	9	9	19

Indel size	Genotype		Number of genotype calls
12	15	15	4
12	15	0	2
12	6	6	1
12	21	0	1
9	9	9	23
9	6	6	17
9	12	12	4
9	0	0	3
9	15	0	2
9	21	0	1
6	0	0	23
6	6	6	18
6	9	9	5
6	9	0	2
6	15	0	1
6	12	0	1
3	0	0	46
3	6	6	3
3	9	0	1
-3	0	0	42
-3	-6	-6	7
-3	-9	-9	1
-6	0	0	30
-6	-6	-6	17
-6	-9	-9	3
-9	-6	-6	22
-9	-9	-9	21
-9	-12	-12	3
-9	0	0	3
-9	0	-12	1
-12	-9	-9	33
-12	-12	-12	13

Indel size	Genotype		Number of genotype calls
-12	-6	-6	3
-12	-15	-15	1
-15	-12	-12	27
-15	-15	-15	15
-15	-9	-9	6
-15	-6	-6	2
-18	-15	-15	32
-18	-12	-12	12
-18	-18	-18	5
-18	-21	-21	1
-21	-18	-18	21
-21	-15	-15	20
-21	-21	-21	7
-21	-12	-12	2
-24	-21	-21	24
-24	-18	-18	23
-24	-15	-15	2
-24	-24	-24	1
-27	-21	-21	32
-27	-24	-24	12
-27	-18	-18	6
-30	-24	-24	26
-30	-21	-21	18
-30	-27	-27	4
-30	-18	-18	2

Table 2.2: Results from simulations of homozygous indel calls. The first column indicates the size of the simulated homozygous indel, the second and third column is the value of the reported genotype from our model and the fourth column is the number of genotypes reported for that particular indel simulation size (out of 50 for each homozygous simulation).

As shown in table 2.2, our model rarely calls homozygotes heterozygotes (16

out of 1,000 incorrectly called heterozygotes, 1.6%) and when this happens, the incorrect genotype always has a reference call, which is in line with the higher prior probability for heterozygous indels with one reference allele. Out of the

Incorrectly genotyped homozygotes as heterozygotes		
Indel size	Genotype	
18	30	0
18	24	0
15	24	0
15	21	0
12	21	0
12	15	0
12	15	0
9	21	0
9	15	0
9	15	0
6	9	0
6	9	0
6	15	0
6	12	0
3	9	0
-9	0	-12

Table 2.3: Simulations where a homozygous indel was called a heterozygote. The first column indicates the size of the simulated homozygous indel, the second and third column is the value of the reported genotype from our modeling.

calls which we correctly called as homozygous (984), 255 of the calls were of the correct size (25.9%), 475 were within ±3 bp (48.3%), 216 were within ±6 bp (22.0%), 36 were within ±9 bp (3.7%) and 2 were within ±12 bp (0.2%).

2.8.3 Heterozygous with one reference allele

Having tackled the two homozygous scenarios (reference and homozygous indel), the heterozygous simulation with one reference allele is essentially a marriage between the previous two. The same number of sample sequences are generated where one corresponds to the reference length and the other is calculated as described in equation 2.8. The difference being that the reads are simulated from

both samples and then amalgamated and aligned to the reference genotype. In practice, paired end reads are sequenced from one of the two copies at a 50% probability (as described in 2.6.3.1). In order to emulate that, the number of paired end reads were first calculated which yielded the desired c for the reference length. Next, a random number generator was used to assign a value between [0,1] for each of the paired end reads to be simulated. Depending on the value of the generated number, the number of reads coming from a given copy (≤ 0.5 for allele 1 and > 0.5 for allele 2) was determined. Once this was complete, the number of paired end reads for each respective repeat length were simulated and then combined into a single set and aligned to the reference repeat length sequence. As the paired end reads were now drawn from two separate distributions, a distinctive bimodel distribution will be observed in the histogram of MPERS for spanning reads (see figure 2.13). This does, in turn, lower the number of reads being drawn from each copy, diminishing the precision of our calls. But given the variant is of sufficient size, our model is able to detect it. In total, out of the 1,000 simulations (50 simulations at each indel size from [-30,30] in units of three bp), only 382 were called reference (38.2%, 100 of which were ± 3 bp that were all called reference). The detection increases to 47% once the indel increases to an absolute size of 12 bp, and rises further to 96.5% for indels with an absolute value over 20 bp (see table 2.4). One source of error for our predictions is for calling heterozygotes homozygotes. The reasoning behind this is that its difficult to distinguish a homozygote site from a heterozygote the mean of whose two indel sizes is the size of the homozygote. Out of the 618 detected variants, 425 were called homozygous (69%) with almost all calls being within a couple motif lengths of the mid value between the variant and the reference. However, when our model did call the site heterozygote, almost all the putative variants were within a few motif lengths of the true variant size. Furthermore, a distinct bias in power to call deletions over insertions is shown in table 2.4. This discrepancy may be caused by the fact that the expected number of spanning reads is greater for a deletion allele as there is less sequence to map across (as described in section 2.4) yielding more unique positions a mate pair can map to.

Genotyped heterozygotes			
Genotype		Detected	Count
30	0	notdetected	1
30	0	detected	49
27	0	detected	50
24	0	notdetected	2
24	0	detected	48
21	0	notdetected	3
21	0	detected	47
18	0	notdetected	11
18	0	detected	39
15	0	notdetected	15
15	0	detected	35
12	0	notdetected	27
12	0	detected	23
9	0	notdetected	39
9	0	detected	11
6	0	notdetected	47
6	0	detected	3
3	0	notdetected	50
0	-3	notdetected	50
0	-6	notdetected	46
0	-6	detected	4
0	-9	notdetected	37
0	-9	detected	13
0	-12	notdetected	26
0	-12	detected	24
0	-15	notdetected	13
0	-15	detected	37
0	-18	notdetected	7
0	-18	detected	43
0	-21	notdetected	8
0	-21	detected	42
0	-24	detected	50
0	-27	detected	50
0	-30	detected	50

Table 2.4: Detection counts of simulated heterozygotes. The first two columns indicate the simulated genotype. The third column is the category for whether a variant was detected or not and the fourth column being the count.

2.8.4 Heterozygous with no reference allele

The last and most complicated, the heterozygous genotype where neither copies' length matches the reference length required generating three samples of lengths l, $l+i_1$ and $l+i_2$. From here, the same procedure as in the heterozygous simulation with one reference allele was carried out for the number of paired end reads to be simulated from each copy, but this time, the reads came only from one of the two sequences that contained an indel. The simulated reads were then mapped to the reference sequence, yielding a bimodal distribution of spanning paired end reads as seen in figure 2.13. For our simulations, we iterated through

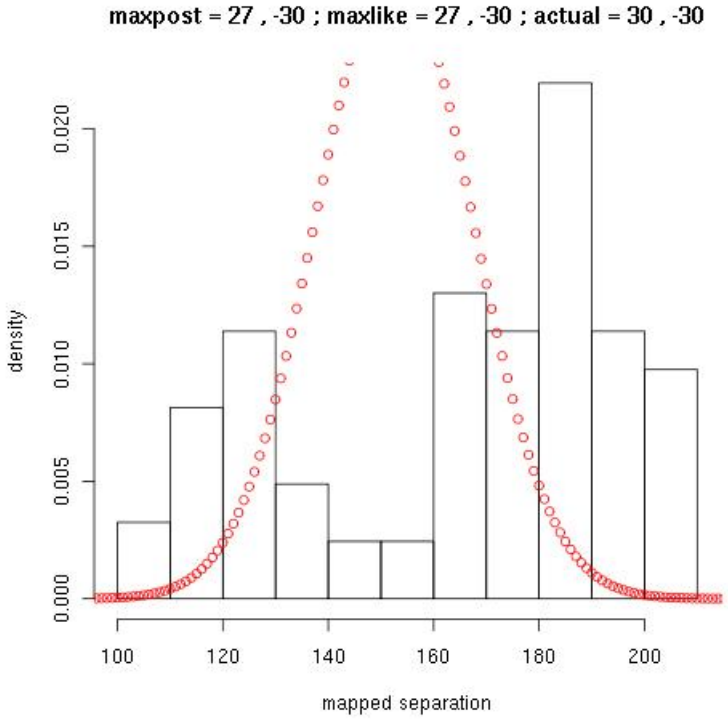


Figure 2.13: Histogram of spanning paired end reads across an individual whose two copies differ in length from the reference graphed against the MPERS distribution for a reference length genotype. This sample contains an insertion of 30 bp and a deletion of 30 bp. It is quite obvious that the number of spanning paired end reads is larger for the deletion allele (peak at right) compared to the insertion allele (peak at left). This is caused by the fact that as the copy lengths are quite different in size (60 bp), the allele containing the deletion is much shorter and therefore has a higher probability of more paired end reads spanning its locus.

every possible heterozygote pair from [-30,30], in units of three bp, excluding the reference allele (described above). In all, we generated 9,500 simulations, which equates to 50 simulations for each genotype. In total, 2,374 were called reference (25%) which is higher than that of the heterozygote simulations with one reference allele. The cause of this increase in reference calls is due to the same problem as described above with the mean of the two variants being called homozygous. In this simulation, indel pairs whose values essentially cancel one another out – an insertion of 12 bp and a deletion of 12 bp – will many times be called reference. Furthermore, as we strongly penalize heterozygotes, many loci with alleles only a couple motifs or less apart will sometimes be called reference because the separation of distributions isn't enough to produce a large enough signal to overcome the prior cost. However, when a variant is detected, its true allele values are within a few motifs – unless pushed into a homozygous configuration the mean of the two variants (4,391 out of 7,126, 62%).

2.9 Results on real data

We have developed a method for inferring the genotype of a STR locus in a diploid sample based on short paired end read sequencing data (see section 2.6) implemented in the software package STRYPE (see section 2.7). For each repeat locus in the reference genome, we assume that a sample has two copies of the repeat locus of lengths $l+i_1$ and $l+i_2$ bp. Based on the short paired end sequence data from the sample, we estimate what sizes of indels i_1 and i_2 in the two copies of the repeat locus relative to the reference genome maximize the *a posteriori* probability found using Bayes' theorem (see equation 2.4), including the case $i_1 = i_2 = 0$. Here we evaluate the use of STRYPE to assay a full genome's worth of tandem repeats for individuals sequenced by both a single and multiple libraries.

2.9.1 Inferring genotypes at repeat loci in individual NA12878

To test the efficacy of our method on a real data, a full assay of all triplet repeat loci ($k = 3$) in NA12878 was conducted. As described earlier in table 1.1 in chap-

ter 1, TRF identified 86,435 triplet repeat loci in the human genome. However, we decided to limit our exploration solely to the autosomes which brought the count of loci down to 80,868 which ranged in length from 15-3925 bp (mean 27.7 bp, median 21 bp). The accuracy of our method depends on the number of spanning paired end reads that are observed across a triplet repeat locus; therefore we only considered loci at which we had $\geq$10 spanning paired end reads. This cutoff was arbitrarily chosen as it was obvious that having only a few spanning paired end reads yielded almost no information – and had we required too many, we would have dismissed a large number of loci (9,113 were dismissed from the 75,688 which had at least one spanning paired end read, figure 2.14). Our prior should remove any further loci that do not contain sufficient information to make a non-reference call. At the 66,575 triplet STR sites with at least ten spanning paired end reads, our method made the following calls: 62,418 reference loci, 3,043 homozygous indel loci, 1,040 heterozygous reference loci (one reference allele and one non-refernce allele) and 74 non-reference (two different non-reference alleles) heterozygous indel loci.

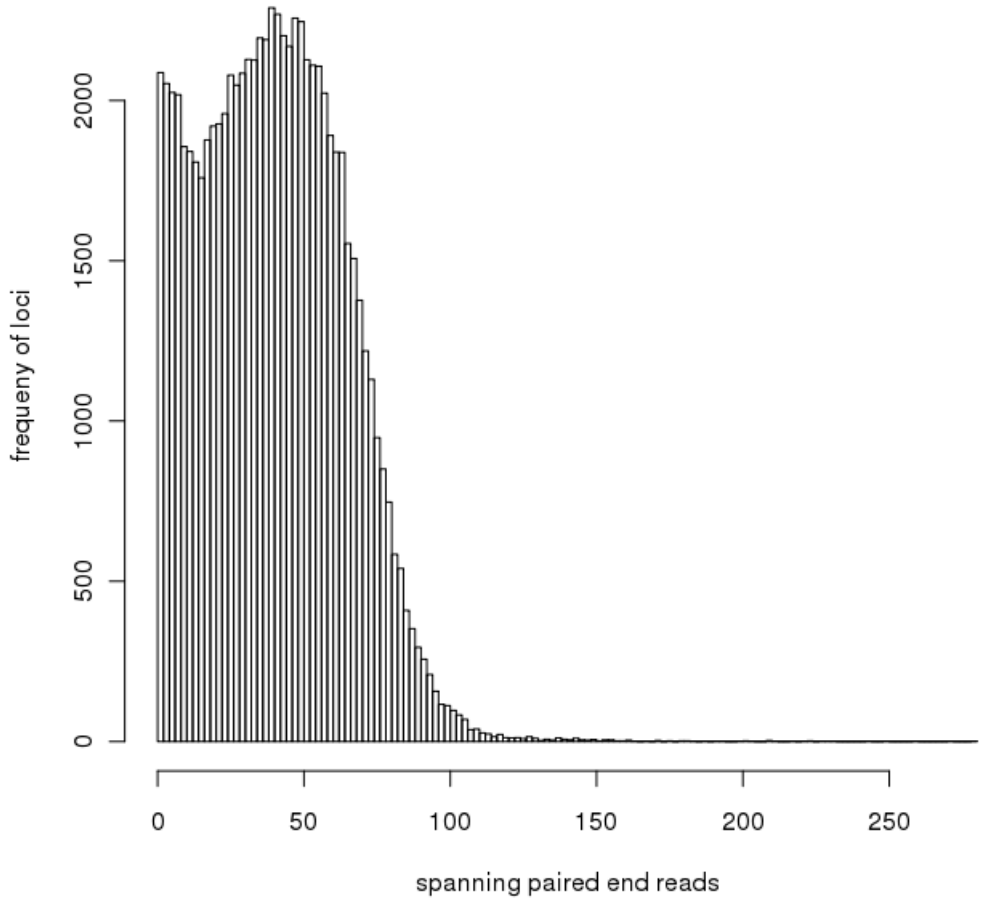


Figure 2.14: Histogram for loci containing a given number of spanning paired end reads for every triplet repeat loci in individual NA12878. The smaller libraries' number of spanning paired end reads will diminish much quicker than the larger fragment libraries and this could explain why the histogram has two peaks as this histogram does not take the size of the tandem repeat in the reference into consideration, only the number of spanning paired end reads.

2.9.2 Accuracy in inferring genotypes at repeat loci

2.9.2.1 Validation data from capillary and 454 alignments

As part of the pilot project for the 1000 Genomes Project, individual NA12878 was sequenced to approximately 22.5x depth using the Illumina sequencer (Consortium [2010]). The same DNA was also sequenced using the 454 sequencer and capillary sequence to approximately 12.8x and 0.5x depth, respectively. The 454

and capillary reads are relatively long (mean 276 and 722 bp), compared to the Illumina reads (mean 37 bp). Because of their length, the 454 and capillary reads are long enough that it is possible to accurately infer some haplotypes (capillary) and genotypes (454) at STR loci by making read-to-genome alignments.

We used automated analysis based on the capillary reads to generate a candidate set of independently confirmed STR indel sites. We then manually inspected 454 alignments at a subset of these sites using the tview alignment tool in samtools (Li et al. [2009]) to produce a truth set for assessing our method's accuracy. Because of the low capillary depth of 0.5x, most loci had only a single allele typed by the capillary reads. In total the capillary analysis called 64 sites with two distinct alleles, 8,463 sites with one called allele matching the reference, and 783 with one called indel allele. The candidate set was composed of all 64 heterozygous calls, plus 114 reference called sites with $\geq$4 spanning capillary reads and 158 indel sites with $\geq$2 spanning capillary reads (see table 2.5).

Validation table for multiple sequence types in individual NA12878			
Capillary call	Number of candidates	454 call	454 call totals
reference	114	reference homozygous indel heterozygous inconclusive	111 0 2 1
homozygous indel	158	reference homozygous indel heterozygous inconclusive	5 56 52 45
heterozygous	64	reference homozygous indel heterozygous inconclusive	4 3 44 13

Table 2.5: Statistics for validation set for multiple sequence types. The capillary calls were used to identify sites of interest based on the number of reads which covered the tandem repeat loci (as discussed in section 2.5). These sites were then examined by eye with 454 alignments to ascertain the true genotype of the locus. The last column states the breakdown of what genotypes were actually observed by eye using the 454 alignments. Some alignments were not readily resolvable by eye due to 454's rate of sequencing errors, especially around repeat units (Huse et al. [2007]).

After visual inspection of the 454 alignments in tview, we removed any sites where the alignments remained unclear (59 sites in total were removed, table 2.5). Figures 2.16 and 2.15 depict two loci where we are both able and unable, respectively, to make the correct genotype call based on visual inspection.

Figure 2.15: Samtools tview of a 454 alignment for an unambiguously genotyped locus. The locus is of repeat motif CAG between positions 165776248 and 165776284 on chromosome 1. From the automated capillary analysis, two separate indels were observed: 3 and -15 bp. When looking at this alignment, it is clear that some of the reads are missing 15 bp of sequence (denoted by blue dash at right) while the others contain an additional 3 bp (yellow and red dashes at right with the inserted motif appearing at the start and end of the repeat, respectively).

Figure 2.16: Samtools tview of a 454 alignment for an inconclusive genotyped locus. This locus is of repeat motif CAA between positions 61801605 and 61801620 on chromosome 1. From the automated capillary analysis, a single indel of -1 bp was called from 2 reads that extended across the locus. Towards the end of the repeat, it appears there is a series of sequencing errors brought on by the poly-A chain that limits our ability to correctly genotype this locus using 454 reads. Because of this, this locus was removed from our analysis.

In the end, we were left with a validation set of 277 calls: 120 homozygous reference ($i_1 = 0$, $i_2 = 0$), 59 homozygous indels ($i_1 = i_2$, $i_1 \neq 0$), and 98 heterozygous loci ($i_1 \neq i_2$). The lower limit of four spanning reads was to ensure that if we only inferred one allele at a particular locus based on the 454 data, it is unlikely that NA12878 is actually heterozygous at the locus and that we simply have not observed the other allele. The probability of observing only a single copy four times and never the other copy is $\frac{1}{2}^4 = \frac{1}{16}$. From the validation set of 277 call sites, our method was able to infer the genotypes at 246 loci (the other loci having too few spanning paired end reads): 117 homozygous reference genotypes, 52 homozygous indel genotypes and 77 heterozygous genotypes, 69 of which contained a reference allele length. Overall, STRYPE's sensitivity to detect indels was good. Figure 2.17 shows the distribution of allele sizes for true indels when STRYPE called a reference genotype (false negative calls).

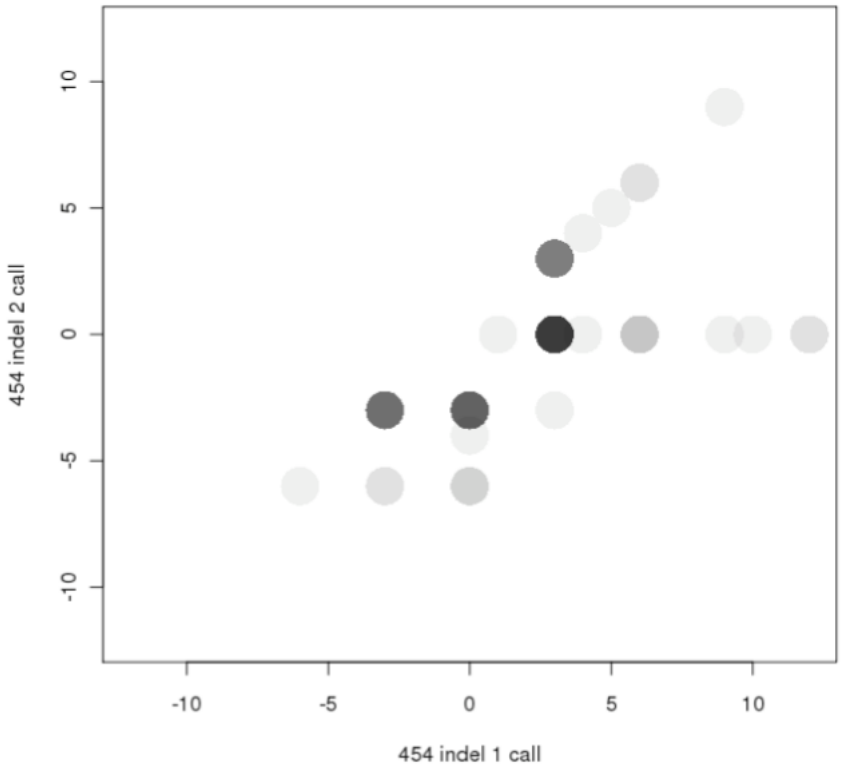


Figure 2.17: Plot of the 454 indel genotypes when our method called a reference genotype, {0, 0}. Almost all these genotypes' repeat lengths are within ±3 bp of the reference length (86%) as demarcated by the four dark black dots around the reference call. As the absolute difference between the reference locus's repeat length and the individual's allele's repeat length increases, so does the power of our method to detect these variants, which explains why fewer and fewer calls appear as you move away from the reference as shown by the light colored, sparsely placed dots.

2.9.2.2 Accuracy at homozygous reference loci

Of the 117 loci inferred to have homozygous reference genotypes in NA12878 based on the 454 data, our method correctly inferred 114 (97.4%) to also be homozygous reference. However, it erroneously inferred one (0.9%) of the homozygous reference loci as a homozygous indel locus and two (1.7%) to be heterozygous (both containing one reference length allele). We were able to fix this by looking at the odds ratios we previously calculated and determining a cutoff which minimized the false discovery rate while not causing too high a number of

true calls to be called reference (see sections 2.9.2.3 and 2.9.2.4). In our model, the calls non-reference calls we are most certain of are those with a high odds ratio between the genotype call made compared to the reference. When we discarded indel calls at which the log odds ratio was weak, ≤ 1, two of the three false positives were removed. By filtering using the odds ratio, it is possible to discard almost all the false positive calls while retaining a large majority of the true calls, 37/44 (84%, see below for discussion of true calls sections 2.9.2.3 and 2.9.2.4). We therefore recommended using this filter because minimizing the number of false positives typically outweighs the loss in number of true indels.

2.9.2.3 Accuracy at homozygous indel loci

Using the 454 sequence data, we inferred that 52 loci have homozygous non-reference indel genotypes. Figure 2.18 illustrates the relationship between what the observed true genotype is – as found by the 454 sequence – compared to what our method calls at these loci. Approximately half the loci (25) had homozygous indels of size of ± 3, only one of which was called as non-reference by our method, indicating that there is insufficient power with these libraries/coverage for our method to distinguish an offset of 3 bp from the reference genotype call. Of the remaining 27 loci, our method calls 21 (78%) as non-reference homozygotes. All but one of our method's calls was within 6 bp or less of the 454 call (the exception being of size +9 bp called as reference), and 5 of 9 sites with absolute indel size 6 were called non-reference. Of the 21 non-reference calls made by our method, all are in the correct direction (no insertions called deletions and vice versa), 8 (38%) are called with the correct size, 11 with absolute difference 3 bp (52%), and the remaining 2 with absolute difference 6 bp (10%). The mean absolute error was a little larger for homozygous insertion (3.3 bp) compared to homozygous deletion (2.7 bp) genotypes.

2.9.2.4 Accuracy at heterozygous loci

Heterozygous genotypes are more difficult to correctly genotype as the number of spanning paired end reads for each copy is approximately half of what it would be compared to a homozygous site. It is also much more difficult to distinguish

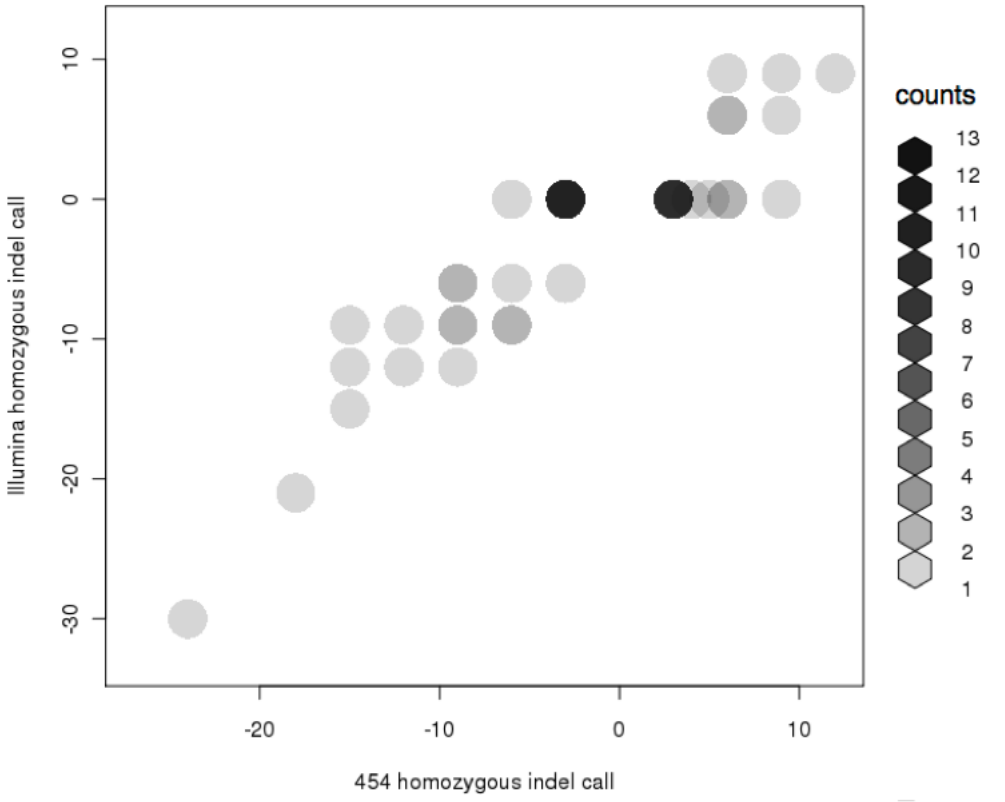


Figure 2.18: Comparison of true homozygous indel genotypes as called from 454 sequence to that of our method's calls at these loci. The diagonal $x = y$ represents our methods calls being exactly on the true genotype call with any deviations from this line an error. Most all calls are within 6 bp from this line. The horizontal value $y = 0$ is representative of loci where there is not enough paired end read information to call a non-reference call. This is where the only outlier of +9 bp lies.

a homozygote site from a heterozygote the mean of whose two indel sizes is the size of the homozygote (as discussed in 2.8.4). To test the efficacy of our model to call heterozygotes, we looked at sites which contained two distinct copies at a locus from our 454 assessment. Based on this 454 data, we inferred 77 loci to have heterozygous genotypes with at least 10 spanning paired end reads: 69 with one reference allele and one non-reference allele and 8 with two different non-reference alleles. Again, true heterozygotes with maximal indel sizes of ± 3 were not called. Out of the 77 loci, 3 (4%) sites were called exactly using our

method. This is lower than for loci with homozygous reference (97.4%) or homozygous indel genotypes (15%). Among loci inferred from 454 data to have heterozygous genotypes with one reference allele (69 sites), 8 were also inferred by our method to be heterozygous with one reference allele. Amongst these, our method correctly inferred the non-reference allele (indel) size at 3 (38%) loci and at 7 loci (88%) the calls were within ±3 bp. At one site the allele difference was 6 bp (our method's call of 12, 454 call of 6). Considering all sites with heterozygous genotypes, our method called 80% of the alleles to within ±3 bp. Figure 2.19 shows how our method performs at the heterozygous loci.

Ultimately, the most telling statistic is the comparison of haploid calls between what the true copies' lengths were as called by 454 sequence to what our method reported. When comparing proximal size alleles in each of the haplotypes of our calls to the true lengths, it is clear that our method is rarely off by more than ±6 bp. What is meant by 'proximal size alleles' is when matching the two copies' lengths to the true lengths, we look for pairings which minimize the absolute difference between the two sets and in case of a tie, take exact matches preferentially. For example, had our method called a locus of genotype {-6,-3} and the true genotype was {0,-3}, then we would match haploids of 0, -6 and -3, -3 as opposed to -3, -6 and 0,-3.

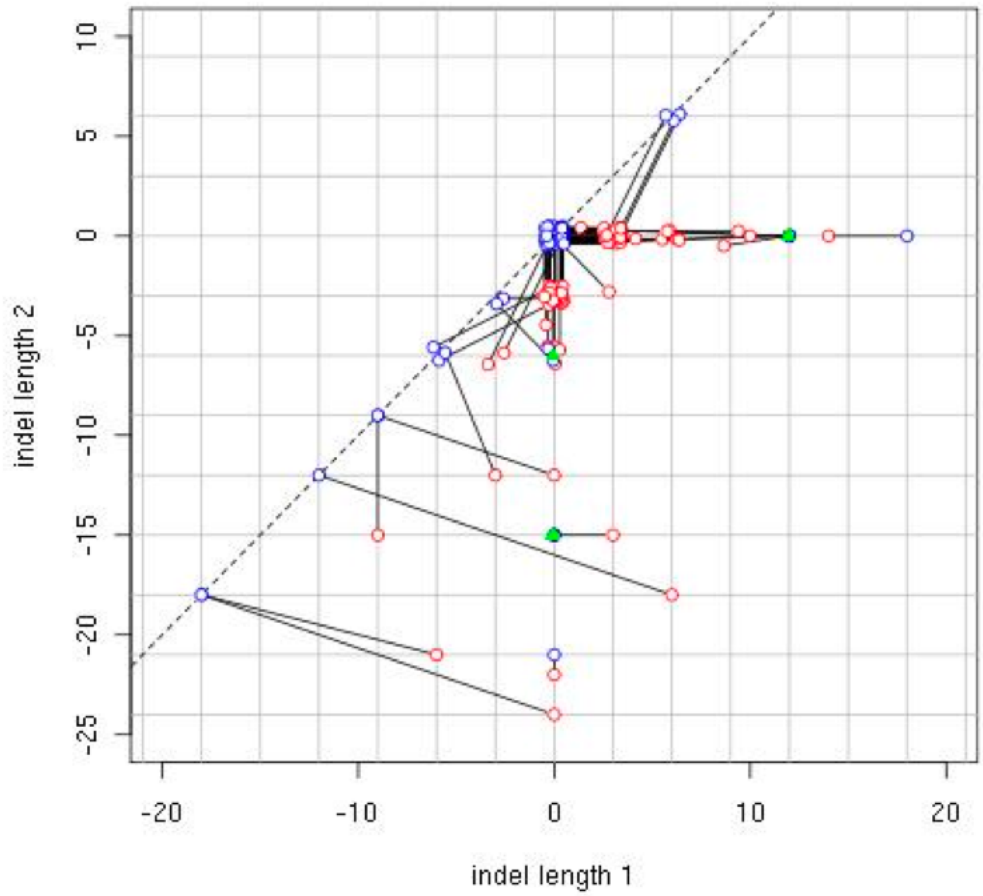


Figure 2.19: Join plot comparison of actual genotype (red dot) compared to the genotype called by our method (blue dot). The dotted diagonal line represents the homozygous genotype while the solid black lines illustrate the difference in genotype calls between the truth and out method's call. Ideally, the shorter the line, the more accurate the call. Horizontal and vertical lines are also significant as they denote that one allele length is called correctly. The three green triangles denote the genotypes where our method accurately called the true genotype. It should be noted that many of the calls overlapped which obfuscated the true number of loci conferring to each genotype call. To alleviate this problem, a random jitter in the range of [-0.5,0.5] was added to all calls that were of a distance of no more than ten units from the reference genotype call. The distance of 10 was chosen as the majority of overlapping calls fell within this range as it represents the smaller, more abundant indels in the genome. Distance is calculated simply as: $distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$, where $x_1 = y_1 = 0$

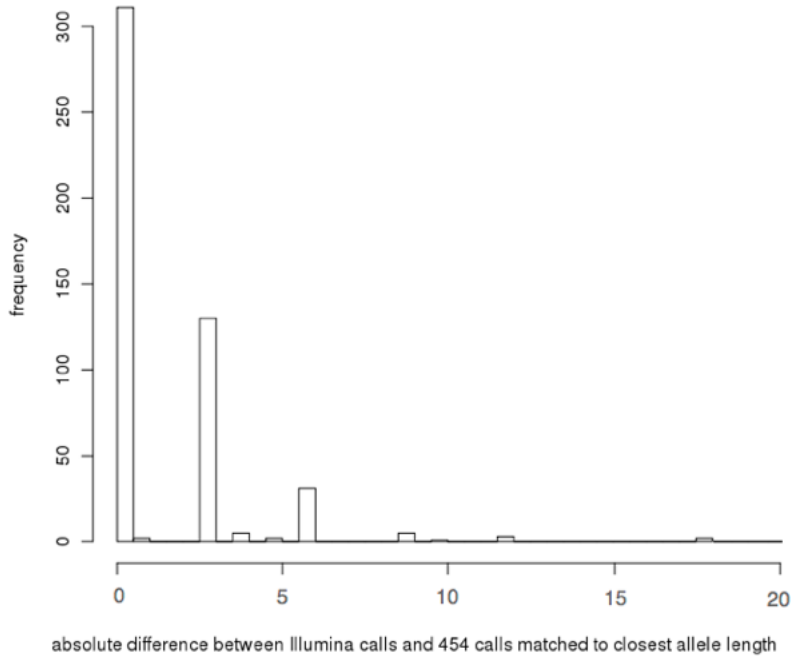


Figure 2.20: Histogram of differences in proximal allele lengths between genotype calls made by 454 and our method. More than half (63%) of all of our method's allele lengths match exactly the allele length inferred from 454 reads. When the threshold is raised to ±3 bp, the percentage raises further to 90%.

2.9.3 Comparison with MoDIL

Of all the other methods which use short paired end reads to detect indels, MoDIL (Lee et al. [2009]) is the closest to our model as it analyses the MPERS distribution of spanning read pairs to infer indels. However, MoDIL is not specifically designed to infer indels in STR loci but indels across the entire genome. This has the added benefit of being robust in calling indels, but lacks in the precision we hope to achieve.

We described how MoDIL works in chapter 1 (see section 1.6.2).

Like our method, MoDIL can infer both homozygous and heterozygous indel genotypes. However, an advantage of our method is that it calculates a confidence score: the odds ratio between the genotype call made and the homozygous

reference genotype (see equation 2.6). Furthermore, while MoDIL assumes that an individual was sequenced from one fragment library, our method can combine data from multiple libraries that differ in the mean, variance and shape of the fragment length distribution. Because of this, it makes a direct comparison of our model's calls to MoDIL's calls for individual NA12878 extremely difficult. The detection power of MoDIL is reported to be 38% for small indels of 10-14 bp and 71% for indels of 15-19 bp (Lee et al. [2009]) on a deeper sequenced sample (NA18507), whereas our method detects 34% (44/129) of indels of any size, 23% (25/107) of variants less than 10 bp, 86% (19/22) of variants greater than or equal to 10 bp in our NA12878 assessment – which is sequenced from multiple libraries to a lower depth. Had we used a single, well behaved library – a library whose distribution is closely inline with a tightly distributed (STD$\leq$10%) Gaussian – which was sequenced to a high depth, we believe STRYPE's proportion of calls would increase further past MoDIL's resolution.

As MoDIL was not designed specifically to use multiple libraries, we were unsure of what MoDIL's efficacy would be by combining the libraries of NA12878. However, through correspondences with MoDILs author, we were told that it was acceptable to add all the libraries together, thus increasing the effective coverage. However, further discussion with MoDILs author suggested that due to the size of the indels we were focusing on (ranging from [-15,15] bp), and in combination with NA12878's libraries standard deviations (ranging from 9.1 to 144.6 bp), MoDIL would be unable to make any calls for indels of this magnitude – even if the paired end reads had been sequenced from the same sample. As outlined in MoDILs supplementary methods, it had a recall rate for indels larger than 10 bp of roughly 0.5 from a single, tightly distributed (STD <10% mean) simulated library of coverages between 5 and 100x (much greater than NA12878's libraries coverage).

Ultimately, to test our belief that MoDIL was unable to make any calls, we ran NA12878's chromosome 11s paired end reads with MoDIL. Since MoDIL cannot specifically target a region, we were forced to run the entire chromosome, which

took magnitudes more time to run than STRYPE. In the end, MoDIL was unable to locate any clusters signifying a structural variation within chromosome 11. Because of this, no indel calls were reported.

2.10 Discussion

We have developed a novel method which uses short paired end read sequencing data to infer the genotype (repeat length) of the two copies of a STR locus in a diploid individual. Our method estimates the lengths of the two indels – one in each of the two copies of the individuals locus – and then calculates the odds ratio between the genotype that maximizes the posterior probability and the reference genotype posterior.

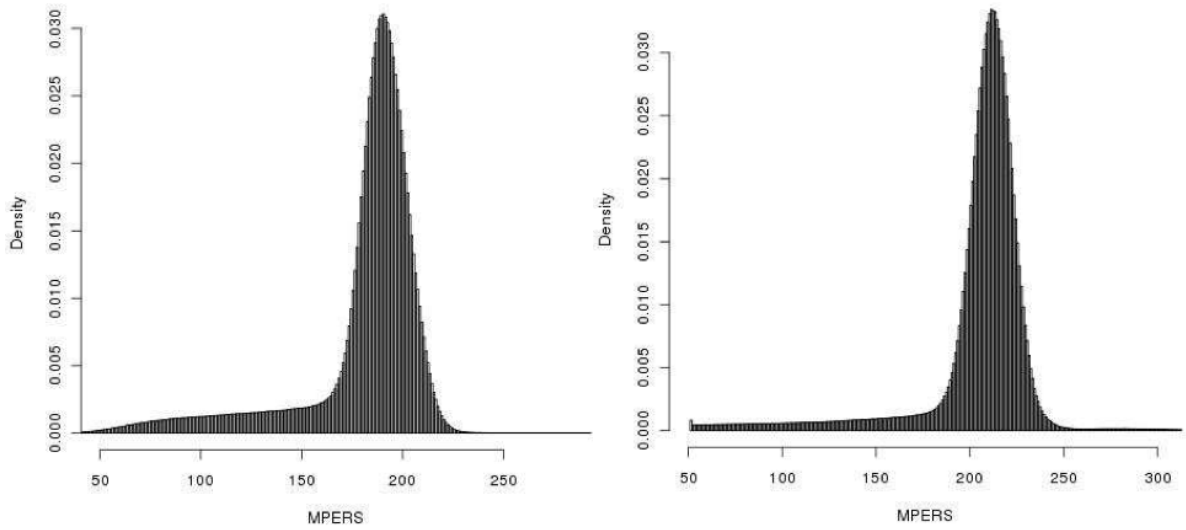
2.10.1 Specific adaptations for detecting indels in STR loci

We have assessed the accuracy of our method by inferring the genotypes at triplet repeat loci in individual NA12878 based on short read paired end data. The accuracy of our method depends on the tightness of the fragment size distributions in each library of NA12878, as well as its overall sequence depth. With an overall average MPERS of 200 bp and standard deviation of 59 bp (see table 2.1), NA12878 is representative of an individual sequenced from multiple semi-well behaved libraries – libraries whose distributions are not as tightly distributed and symmetric as the Gaussian. With the libraries having a combined depth of 22.5x, our method can discover a majority of variation $\geq$6 bp with few to no false positives.

Overall, our method correctly inferred the genotype at 63% of all triplet repeat alleles and 90% of all triplet repeat alleles within $\pm$3 bp (see figure 2.20). One limitation of our method is that it requires a reasonable number of spanning paired end reads ($\geq$10) to infer the genotype at a repeat locus. While NA12878 was sequenced to a depth of 22.5x, for a variety of reasons some genomic regions had a much lower physical coverage. We found at least one spanning read pair at 77,165 (95%) of the 80,868 triplet repeat loci located in autosomes identified

by TRF, and ≥10 spanning read pairs at 66,575 (82%) loci. Reasons for not having enough spanning read pairs include base composition bias of the sequencing libraries, non-uniqueness in the flanking sequence and the repeat being too long. The mean fragment length per library for many of the NA12878 libraries is above 200 bp (see table 2.1), so we could, with sufficient depth, be able to infer genotypes for loci of up to 200 bp. This includes most triplet repeat loci since less than 1% of triplet repeats are longer than 200 bp in length.

Our method calls more deletion alleles (5282) than insertion alleles (1992). One reason for this is that we lose power to call large insertions in long STRs because these variants can result in total lengths longer than the paired end separation. However, almost all STRs detected with insertions have lengths that are shorter than the MPERS distribution, therefore the primary reason for the imbalance is that many of the libraries for NA12878 have a heavy left-tail in the fragment size distribution (see figure 2.21). As leftward shifts of the MPERS distribution for paired end reads spanning a locus are used by our method to infer an insertion, this reduces our power to detect these events. Generating libraries with a tighter, more symmetric distribution of fragment lengths will alleviate this problem.



(a) Distribution of the MPERS for library g1k-sc-NA12878-CEU-2

(b) Distribution of the MPERS for library Solexa-5460

Figure 2.21: Distribution of the MPERS for two separate libraries for sequenced individual NA12878. A noticeable heavy left-sided tail can be observed which lessens the statistical power for calling insertions.

2.11 Conclusion

In conclusion, we have developed a novel method for inferring genotypes in STR loci based on short paired end read data and have identified 4,157 loci with non-reference STR variants in NA12878 with a low false positive rate. This data set and method helps give a more complete picture of genetic variation based on whole genome next generation sequence data, and will aid in studies of STR mutation and evolution.

Chapter 3

Factors influencing polymorphism in short tandem repeats

Collaboration note *This chapter contains work performed in collaboration with Dr. Avril Coghlan and Dag Lyberg. Avril assisted in curating a list of triplet repeat positions in the human genome which contained the locus's repeat motif and motif family. Dag assisted in curating a list of transcript sites from ENCODE.*

The hypermutability of STRs makes them of great interest to geneticists. Many smaller surveys have been conducted to ascertain the mutation rate of short tandem repeats (Lai and Sun [2003], Whittaker et al. [2003], Brinkmann et al. [1998], Ananda et al. [2011]). These studies have focused on a small set of specific loci in the human genome (Brinkmann et al. [1998]; Weber and Wong [1993]) due to the complexities of typing short tandem repeats (as discussed in chapter 2).

Past research has sought to understand their evolution over time (Calafell et al. [1998]) as well as use STRs as markers for forensic analysis (Kasai et al. [1990]; Urquhart et al. [1994]; Lygo et al. [1994]; Ruitberg et al. [2001]). As of the writing of this work, there has been no genome wide assay of short tandem repeats that we are aware of. A genome wide assay of STRs would have the power to elucidate what factors in STRs increase the chance of observing a variant at a locus. Some of the proposed factors include the composition of the repeat motif, the purity of the repeat in the reference genome, the length of the repeat in the reference

genome, the GC content of the repeat and proximal sequence and whether the STR resides within a transcript. There has been past research that looked into understanding how some of these factors affect mutation rate (Xu et al. [2005]), but nothing on a large, genome wide scale. This is due mostly in part to the fact that past sequencing of STRs is both costly and slow (Sprecher et al. [1996]), which has precluded a large, genome wide assay. However, due to the advent of next generation sequencing technology, we are now able to explore these loci on a massive scale.

Building upon our method of genotyping STRs using spanning paired end reads (see chapter 2), we plan to understand what factors in a STR increase or decrease the chance of observing a variant at that locus.

3.1 Sources of sequence

To increase the total number of variants found, and therefore the power of our analysis, we ran STRYPE on three trio data sets which met the requirements of being sequenced to a high coverage with Illumina paired end reads. A trio data set derives from a nuclear family composed of each parent and a single child. Two of the trios were from the 1000 Genomes Pilot Project (Consortium [2010]) which consisted of families from the CEU and YRI population, and the third was sequenced by Illumina – also from YRI population HapMap samples.

3.1.1 1000 Genomes pilot trios

The sequence data for both 1000 Genomes Project families is publicly available and can be downloaded from ftp://ftp.1000genomes.ebi.ac.uk/ . These were mapped using the BWA alignment tool as part of the 1000 Genomes pilot project.

3.1.1.1 Sequencing statistics

A summary of the libraries' statistics from the 1000 Genomes trio pilot set is shown in table 3.1.

Library statistics for 1000 Genomes trio pilot data				
Population	Individual	Library	Bases	Coverage
CEU	NA12891	Solexa-6407	7817400156	2.6
		Solexa-3625	43934509439	14.6
		g1k-sc-NA12891-CEU-2	21897329228	7.3
		g1k-sc-NA12891-CEU-1	15129837000	5.0
		totals	88779075823	29.6
	NA12878	g1k-sc-NA12878-WG-1	19327027164	6.4
		Solexa-3630	14717717437	4.9
		g1k-sc-NA12878-CEU-1	12546297144	4.2
		NA12878.1	10463534460	3.5
		g1k-sc-NA12878-CEU-2	6012622836	2.0
		Solexa-5460	4443002700	1.5
		totals	67510201741	22.5
	NA12892	g1k-sc-NA12892-CEU-1	15254665056	5.1
		g1k-sc-NA12892-CEU-2	21865659579	7.3
		Solexa-3594	31658274363	10.6
		Solexa-5455	11074558755	3.7
		totals	79853157753	26.6
YRI	NA19238	2675169269	17346838500	5.8
		QRAAADHAAPE	702666135	0.2
		2485373691	34983913124	11.7
		QRAAADCAAPE	2352597354	0.8
		totals	55386015113	18.5
	NA19240	2675080346	26442703184	8.8
		QRAACDJAAPE	195022575	0.1
		QRAACDEAAPE	8315238204	2.8
		2485441832	50960025784	17.0
		CT1898	22975401315	7.7
		totals	108888391062	36.3

Population	Individual	Library	Bases	Coverage
YRI	NA19239	QRAABDDAAPE	10045560105	3.3
		QRAABDHAAPE	459880560	0.2
		2485443314	37182509292	12.4
		2675080202	30382984800	10.1
		totals	78070934757	26.0

Table 3.1: Mapped bases and corresponding coverage for the two trios in 1000 Genomes pilot project. The first column indicates the population from which the individual (column 2) was sequenced from. The third column indicates the sequenced library and the fourth and fifth column indicate the number of bases sequenced and effective base coverage, respectively, for that library.

3.1.2 Illumina Trio

The sequence data for the Illumina trio is publicly available and can be downloaded from http://www.ncbi.nlm.nih.gov/sra with identifiers SRA009225 (NA18506), SRA000271 (NA18507) and SRA009347 (NA18508). Each of these individuals' libraries were mapped using the BWA alignment tool as part of the Illumina sequencing study. Table 3.2 lists the the libraries from which each individual was sequenced and its corresponding coverage.

Library statistics for Illumina trio data			
Individual	Library	Bases	Coverage
NA18506	CT1696	126419574701	42.140
NA18507	CT1194	125394885034	41.798
NA18508	CT1704	121122865300	40.374

Table 3.2: Mapped bases and corresponding coverage for the Illumina trio data set.

3.2 MPERS distributions

Figure 3.1 shows the distributions of libraries coming from the nine individuals in our data set.

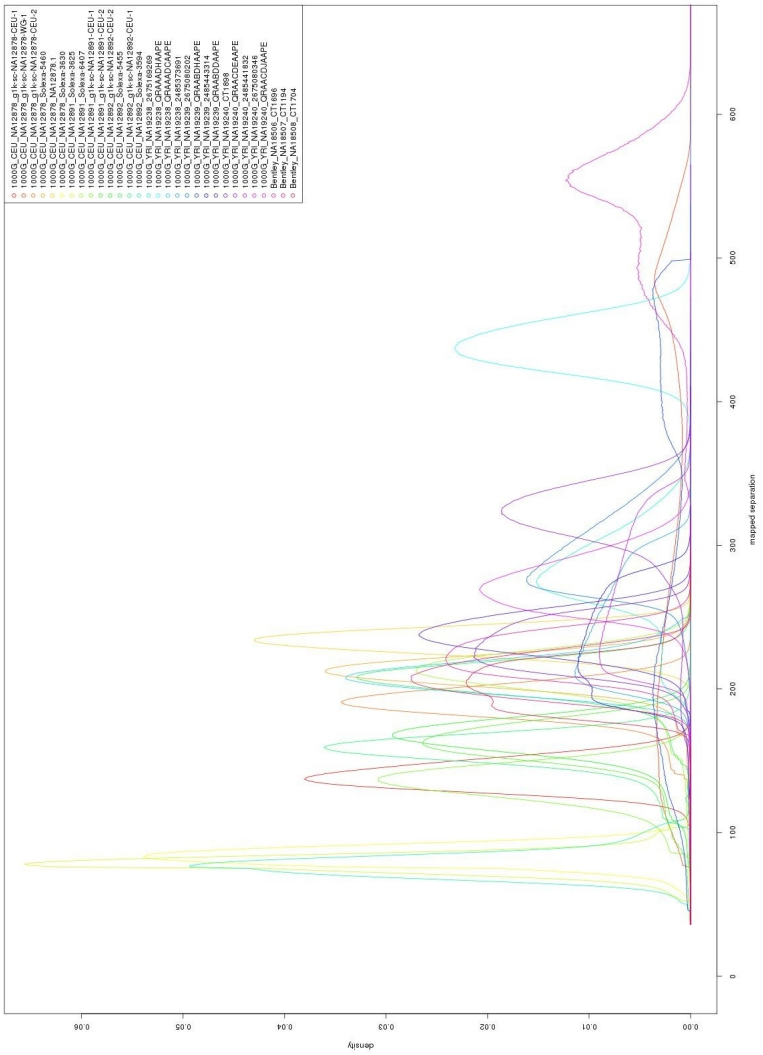


Figure 3.1: Distributions of each library in the nine individuals from the three trios data set. Made up of thirty libraries, the range and shape of each library is unique. The libraries which yield more information for our analysis are those that are tightly distributed around the fragment size library (the peak of the curve, such as those around 80, 150 and 250 bp). The less sharp peaks – as well as those with heavy tails – yield less information from which we can use to genotype STR loci.

Aside from the mean and standard deviation of each library (which sometimes can be misleading), we looked at two statistics that might give us a better sense of how well behaved each libraries' distribution of MPERS really are; skewness

and kurtosis.

Knowing whether a library is symmetric or not is important if we are to understand why one form of indels is being called over the other (as discussed in chapter 2). When a library's distribution is heavy tailed, the sensitivity to call indels that correspond to MPERS shifts in the direction of the heavy tail decreases. Also, a more gradual decline in the density of MPERS as you move away from the mean adds noise to our system when calling indels in that direction. By knowing the skewness of our distributions, we have a better idea of any underlying biases in calling insertions or deletions.

The skewness, γ_1, of each library (which is the third standardized moment) is calculated as

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu^3}{\sigma^3}$$

where μ_3 is third moment about the mean and σ is the standard deviation. From this formula, we were able to calculate the sample skewness of each library from n values (where n is the number MPERS in a library) as

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}} \tag{3.1}$$

where $\bar{x}$ is the sample mean, m_3 is the sample third central moment, and m_2 is the second central moment (sample variance). To elucidate the correlation of moments, the denominator in equation 3.1 was simplified so that skewness was calculated in terms of the ratio of the third cumulant m_3 and the second cumulant, m_2.

As a final statistic, we calculated the kurtosis of each library to get a sense of how peaked our data was around the mean. A higher value for kurtosis meant that more of the variance of the data is a result of extreme outliers as opposed to moderately sized deviations. Explicitly, kurtosis is the standardized fourth

moment and is defined as

$$\beta_2 = \frac{\mu_4}{\sigma^4},$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. This gives rise to the more commonly referred to expression that is defined as the fourth cumulant divided by the square of the second (variance squared) minus 3.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

The minus 3 is a correction to make the kurtosis of the normal distribution equal zero. Lastly, the sample kurtosis for n values was calculated as

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

Table 3.3 outlines the values of these four statistics; mean, standard deviation, skewness and kurtosis for each of the libraries sequenced from individuals in the three trios.

Library statistics for three trio populations						
Population	Individual	Library	Mean	Std	Skewness	Kurtosis
CEU	NA12878	g1k-sc-NA12878-CEU-1	140.177	10.392	0.155	-0.034
CEU	NA12878	g1k-sc-NA12878-CEU-2	189.372	14.805	-1.110	2.906
CEU	NA12878	g1k-sc-NA12878-WG-1	301.076	144.622	0.296	-1.326
CEU	NA12878	NA12878.1	233.048	9.229	-0.190	0.072
CEU	NA12878	Solexa-3630	84.112	7.788	0.103	1.331
CEU	NA12878	Solexa-5460	210.661	14.574	-1.467	6.764
CEU	NA12891	g1k-sc-NA12891-CEU-1	133.592	15.348	-0.674	1.199
CEU	NA12891	g1k-sc-NA12891-CEU-2	157.120	22.587	-1.566	3.436
CEU	NA12891	Solexa-3625	79.055	7.747	0.329	2.272
CEU	NA12891	Solexa-6407	206.715	23.307	-2.054	6.943
CEU	NA12892	g1k-sc-NA12892-CEU-1	155.015	15.309	-1.099	2.030
CEU	NA12892	g1k-sc-NA12892-CEU-2	163.254	18.798	-1.279	2.774
CEU	NA12892	Solexa-3594	78.128	10.278	1.034	3.371
CEU	NA12892	Solexa-5455	204.256	17.115	-1.539	5.571
YRI	NA19238	2485373691	242.773	32.157	0.219	-0.822
YRI	NA19238	2675169269	276.749	32.648	-0.454	0.234
YRI	NA19238	QRAAADCAAPE	209.670	12.329	0.317	0.338
YRI	NA19238	QRAAADHAAPE	439.481	16.263	0.030	-0.049
YRI	NA19239	2485443314	231.121	30.080	-0.059	-0.563
YRI	NA19239	2675080202	294.645	27.179	0.255	0.091
YRI	NA19239	QRAABDDAAPE	238.279	14.850	-0.266	0.394
YRI	NA19239	QRAABDHAAPE	295.971	125.481	0.069	-1.388
YRI	NA19240	2485441832	263.918	40.037	0.183	-0.755
YRI	NA19240	2675080346	273.335	19.263	0.185	0.167
YRI	NA19240	CT1898	230.439	16.299	-0.065	-0.306
YRI	NA19240	QRAACDEAAPE	309.156	42.723	-2.214	5.809
YRI	NA19240	QRAACDJAAPE	527.394	50.323	-1.052	1.437
Illumina	NA18506	CT1696	222.426	15.779	-0.426	0.482
Illumina	NA18507	CT1194	209.138	13.072	0.046	-0.431
Illumina	NA18508	CT1704	202.089	15.247	0.024	-0.450

Table 3.3: Statistics for individuals' libraries in the three trio data sets. The first three columns indicate the population, individual and library from which the statistics are coming from, respectively. And the last four columns represent the mean, standard deviation, skewness and kurtosis of each library.

3.3 Detecting indels in short tandem repeats

Using the methods described in chapter 2, we genotyped all triplet repeat loci for each individual in the three trio sequencing data sets. Each individual was typed independently; no information from which family the individual was from was used to force Mendelian segregation at putative variant sites. Altogether, 596,078 sites had ≥ 10 spanning paired ends across the nine individuals, 29,746 had no spanning paired end reads and 101,727 had < 10 spanning paired end reads. From the sites with ≥ 10 spanning paired end reads, STRYPE called 548,141 loci homozygous reference and 47,937 with a variant. The total number of genotype configurations was in line – relative to one another – with what we would expect: 29,904 homozygous indels (the most likely), 14,957 heterozygous with one reference allele (second most likely) and 3,076 heterozygous with no reference allele. A summary of the three trio family call sets is presented in table 3.4.

Variant call statistics for three trio families				
Individual	Sites called	Sites uncalled	≥10 spanning reads	Reference
NA18508	77946	2893	71834	64747
NA19238	77196	3643	60034	58892
NA19239	78299	2540	70538	67017
NA18507	76143	4696	69833	61457
NA12891	77471	3368	56709	52339
NA12878	78309	2530	69196	62835
NA18506	77969	2870	71523	61804
NA12892	75631	5208	50707	48151
NA19240	78841	1998	75704	70899
total	697805	29746	596078	548141
Individual	Variants	Homozygous indels	Heterozygous reference	Heterozygous
NA18508	7087	5055	1790	242
NA19238	1142	954	147	41
NA19239	3521	2586	798	137
NA18507	8376	5617	2450	309
NA12891	4370	1798	2008	564
NA12878	6361	3410	2427	524
NA18506	9719	5786	3073	860
NA12892	2556	1267	1088	201
NA19240	4805	3431	1176	198
total	47937	29904	14957	3076

Table 3.4: Variant calls made in the three trio families.

3.4 Short tandem repeat criteria

Measuring the prevalence of STR variation as a property of its sequence composition and context has been a goal of this research since the initial modeling of variants in a single sample (see chatper 2). The probability of observing a variant at a locus depends on multiple factors. In the following sections, a list of factors which we believe might influence an STR's chance of exhibiting a variant will be discussed and assessed using the calls made from our three trio families data.

3.4.1 STR metrics

To determine the effect a certain factor has on the prevalence of variation across varying STR loci, it is first important to define what exactly we are measuring in a way that yields a clear mechanism for inference. One way of doing this is by setting forth a metric for each factor. A metric is a simple way of ordering a set such that the distance between each value in a set can be directly calculated. The metric itself will take the form of a set of ordered numbers where a higher order number means either an increase or decrease in a factor we are trying to measure. Each factor we wish to measure has its own metric and in turn, its own strengths and weaknesses. A metric will never encapsulate all the information of a system, but does help us order a set of data which we can later analyse to see what effect (if any) a certain factor has on a system. In the sections below, we describe the factors (listed in table 3.5) we believe will have the greatest effect on observing a structural variation at a locus and how each factors's metric was calculated.

Description of factor tags	
Factor tag	Description
family	trio family from which the individuals come from
motif	triplet repeat motif family from which the STR is a part of
purls	longest stretch purity metric
purnew	purity percent match
GCref	percent of GC content in a STR locus
GC100	percent of GC content in a STR locus and up and down stream 100 bp
GConly	percent of GC content up and downstream 100 bs of a STR locus
lenpurnew	length based metric for purity percent match
trans	boolean value whether a STR is located within a transcript
reflen	length of a STR in the reference
spanreads	number of observed spanning read pairs across a STR

Table 3.5: Table of the factor tags used in our modeling and their respective description.

3.4.2 Tandem repeat length in reference (reflen)

A STR's repeat length in the reference was calculated directly from the start and stop positions of the repeat. As described in chapter 2, all STR loci in the human genome were located using Tandem Repeat Finder (TRF) that met a set of criteria that determined whether a stretch of sequence in the reference should be considered a tandem repeat or not. The length (and in turn metric) was calculated as

$$l = z - y + 1$$

where y and z represent the start and end position of the STR in the reference sequence, respectively. This metric is very basic and tells us nothing about the internal composition of the repeat other than its length. The background mutation rate has been estimated to be on the order of 10^{-8} per base for single nucleotide polymorphisms (Drake et al. [1998]) and approximately a magnitude less for length mutations, 10^{-9} (Nachman and Crowell [2000]). Using just this information, it stands to reason that as the length of the STR locus increases, so shall the probability of observing a structural variation.

3.4.3 Tandem repeat motif family (motif)

Repeat motifs are self-repeating stretches of DNA sequence. These repeats can take the form of any repeating permutation of the four bases {A,C,G,T}. Within these permutations, repeats of the same motif length can be grouped together by their sequence similarities. These similar sequence patterns are grouped together in 'families'.

Each motif length will have some number of families; the simplest example are the motif families for the motifs of length one. Within each family, there is also some number of repeat permutations. Each of the permutations in a family must represent correctly ordered sequence matches of the repeat sequence on the forward strand, as well as its reverse complement sequence on the reverse strand.

For example, the motif family AAC would have three permutations on the forward strand (AAC, CAA and ACA) and three permutations on the reverse strand (TTG, GTT and TGT).

In total, there are 10 unique repeat families for repeat motifs of length three bp – which have been summarized in table 3.6.

List of families for motifs of length three		
Motif family	Forward strand	Reverse strand
AAC	AAC, CAA, ACA	TTG, GTT, TGT
AAG	AAG, GAA, AGA	TTC, CTT, TCT
AAT	AAT, TAA, ATA	TTA, ATT, TAT
ACC	ACC, CAC, CCA	TGG, GTG, GGT
ACG	ACG, GAC, CGA	TGC, CTG, GCT
ACT	ACT, TAC, CTA	TGA, ATG, GAT
ATC	ATC, CAT, TCA	TAG, GTA, AGT
ATG	ATG, GAT, TGA	TAC, CTA, ACT
ATT	ATT, TAT, TTA	TAA, ATA, AAT
CCG	CCG, GCC, CGC	GGC, CGG, GCG

Table 3.6: Table of each motif family belonging to the set of motifs whose repeat length is three.

3.4.4 Purity of tandem repeat in reference

The purity of a tandem repeat is defined as the degree of unbroken repeat units of a motif in a STR locus. This score is effected by the number of foreign base pairs (those that do not match the motif) and inserted or deleted sequence that exist within a repeat locus. The larger amount of foreign bases and indels in a locus decreases the level of purity of that repeat. Purity is an important metric to scrutinize as the purity of sequence in a tandem repeat has been shown to increase the variability at a repeat locus (Legendre et al. [2007]). Many metrics have been proposed in regards to repeat purity. In the following section, we shall discuss three metrics we used to categorize the purity of each tandem repeat.

3.4.4.1 Longest pure stretch (purls)

The longest pure stretch of a STR is the length of the longest subsequence within a repeat locus that goes unbroken by a foreign base either through substitution or addition/removal of a base(s). For example, in the sequence AACAACAACGAACAA, the subsequence AACAACAAC (which is comprised of three full repeat units) is the longest stretch with length 9 bp. Our longest stretch metric does allow for the first and last repeat to be truncated. The longest stretch for repeat sequence TTGTTGTAGTTG would be TTGTTGT, where the two bases TG are removed from the last repeat.

3.4.4.2 Percent match (purnew)

Aside from the longest pure stretch which only measures a subsequence in a STR locus, percent match measures the overall adherence to the motif unit across the locus. This metric gives us a better idea of the overall purity of a repeat locus.

For our analysis, we devised two related metrics to measure the percent match a tandem repeat had to its given repeat motif. The first, purnew, is the overall adherence of a STRs sequence to its repeat motif. This algorithm looks at each subsequence of length of the motif and determines if it matches the overall consensus motif pattern. The algorithm calculates the proportion of start positions in a tandem repeat locus whose subsequent sequence matches the family of motifs a repeat locus is attributed to. It should be noted, however, that it only gives a positive score for subsequences that match the motif on the same strand. For instance, the family of motifs AAC would have AAC, ACA and CAA on the forward strand and TTG, GTT and TGT on the reverse. If the motifs of the reverse strand appear on the forward strand, they are considered foreign bases and not scored as fitting the motif pattern. The second metric, lenpurnew, is simply the value of purnew multiplied by the length of the repeat locus in the reference. This in essence scales the percent match value to the repeat length. We believed it was important to have this additional metric associated with the purnew metric because ignoring the length of the STR gives rise to a bias in

shorter STRs having a higher purity metric score than longer STRs. This bias is described later in section 3.5.1.1.

3.4.4.2.1 Percent match algorithm

1. Set $score = 0$
2. Define all possible permutations which match a family of motifs that reside on the same strand
3. Starting at $x = 1$
4. If subsequence $(S_x, ..., S_{x+|M|-1})$ matches a possible permutation defined in 2, $score++$
5. $x++$

   If $x \leq |S| - |m| + 1$

   goto 4

   else

   last
6. Calculate purity as $\frac{score}{|S|-|m|+1}$

The value of the purnew metric was calculated as described above, yielding a value residing between $[0, 1]$. A higher value is indicative of a larger adherence to the motif family and less foreign bases, indels within the locus.

3.4.5 GC content in and around tandem repeat (GCref, GC100 and GConly)

The amount of GC content in and around a STR can have an impact on both the detection and prevalence of observing an indel. GC rich regions have been shown to have an increased prevalence of sequencing errors (Dohm et al. [2008]; Meacham et al. [2011]). These errors would cause the mapping of paired end reads to decrease, thus decreasing the effective coverage of a locus. For our analysis, we considered three GC composition metrics

1. GCref: the fraction of G or C bases in the reference STR sequence
2. GC100: the fraction of G or C bases in the reference STR sequence plus 100 bp up and down stream
3. GConly: the fraction of G or C bases in the 100 bp flanking regions only

3.4.6 Whether a tandem repeat is in a transcript (trans)

The last metric is whether or not the STR resides within a known transcript (both introns and exons). The human genome's transcript start and stop positions were downloaded from the ENCODE project website at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/. In total there are 70,663 transcripts on autosomes in the ENCODE data base. Of these, many were duplicated (exact start and stop positions) which were removed leaving a total of 51,492 transcripts. Further to this, there were many overlapping transcripts. When determining if a STR resided within a transcript, it only needed to be located in one of the overlapping transcripts. We did not distinguish between multiple transcripts for a single STR; a count of one was given no matter the number of transcripts that the STR was situated in. In total, out of the 80,805 triplet repeat sites in the human autosomes, 42,622 resided within a transcript and 38,183 laid outside.

3.5 Results

We approached our analysis of STR factors in two ways: the influence each factor had on observing a non-reference allele, and the effect each factor had on the overall magnitude of the observed indel for both insertions and deletions. To begin, we sought to determine the effects of observing a non-reference allele by using a logistic regression which determined the influence each factor had on observing a non-reference allele at a given locus. In more detail, a logistic regression is used in predicting the probability of the occurrence of an event by fitting the data to a logit function of a logistic curve. For our purposes, we were interested in the logistic regression as it is a generalized linear model (GLM) used in binomial

regressions (discussed below). Like other regressions, the logistic linear regression can make use of several predictor variables (our factors) that may be either numerical (purity, reference length, etc.) or categorical (motif family, trio family, etc.).

To begin, the logistic function is defined as

$$f(z) = \frac{e^z}{e^z + 1} \tag{3.2}$$

where z is some linear relationship between the explanatory variables

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where β_0 is the intercept and $\beta_1, ..., \beta_p$ are the regression coefficients of the explanatory variables $x_1, ..., x_p$, respectively. The variable z in essence is a measure of the total contribution of all the independent variables used in the model. Next, as mentioned previously, this logistic regression is a GLM for the binomial regression. A binomial regression can be described as a series of Bernoulli trials (a series of one of two possible disjoint outcomes). The results of this regression are assumed to be binomially distributed which is fitted as a generalised linear model where the predicted values μ are the probabilities that any single event will result in a success (indel). The likelihood of these predictions μ are given as

$$L(D|\mu) = \prod_{i=1}^{n} \mathbb{I}_{y_i=1}(\mu_i) + \mathbb{I}_{y_i=0}(1 - \mu_i) \tag{3.3}$$

where D represents the response data, $\mathbb{I}_{y_i}$ is the indicator function which takes the value one when an event occurs and zero otherwise. The likelihood function is specified by defining the parameters μ_i as functions of the explanatory variables (in our case the factors). There are many methods of generating the values of μ in systematic ways that allow for interpretation of the model. However, there is a requirement that the model linking the probabilities μ to the explanatory variables should be of a form which only produces values in the range 0 to 1 which we have described above in equation 3.2. It is then only a matter of fitting

the model to the parameter values that maximize the likelihood in equation 3.3.

Next, we looked at the influence each factor has on the magnitude of an indel given an indel is observed. Sites which were called reference by our model were excluded from this analysis. A linear model was used for this analysis as it determined the value each factor had on the overall value of the response variable – in this case the size of the indel. A linear model is a statistical model which models the relation between the observations Y_i (indels) and the independent variables X_{ij} (factors) as

$$Y_i = \beta_0 + \beta_1(X_{i1}) + \cdots + \beta_p(X_{ip}) + \epsilon_i, \qquad i = 1, \ldots, n$$

where β_i are the regression coefficients and ϵ_i is the residual error. The value of β_0 represents the intercept of the linear model while the rest of the regression coefficients represent the amount of influence (equivalent to slope) a factor has in describing the overall system you aim to model; a positive coefficient denotes a positive correlation while a negative coefficient denotes a negative correlation. Assuming the residual errors are normally distributed, the values of these coefficients are estimated by least squares analysis by minimizing the sum of squares function (S), which is defined as

$$S = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1(X_{i1}) - \cdots - \beta_p(X_{ip})\right)^2 .$$

We used the software package R to carry out this analysis (R Development Core Team [2011]).

3.5.1 Modeling of factors

A logistic regression was used to determine the effect each of the 11 factors had on observing a non-reference allele in a STR locus. For ease of computation and modeling, we separated the called genotypes into two alleles and did all the analysis at the level of alleles. This appeared to be the easiest approach and we did not feel it changed the overall inference we could make regarding the the outcome

of our modeling. A summary of the model's output is produced by R, giving the value of each of the coefficients for each of the factors as well as a p-value that indicated the confidence the model had that each of the coefficient values was non-zero. For almost every coefficient calculated in our analysis, the p-value was less than 0.001. Because of this, when we discuss specific coefficients below they will by default have a p-value less than 0.001. In the rare cases where this isn't the case, we shall explicitly state which factors' coefficients are not statistically significant. This is the same for our linear model which we used to determine what effect, if any, a factor has on the magnitude of an observed indel.

To begin, we looked at the reference and non-reference calls for a combined model incorporating all factors listed in table 3.5. However, this produced some surprising results (see figure 3.2), where for example GC100 had a negative coefficient and GConly had a positive coefficient, although those are themselves strongly correlated. Further investigation showed that this correlation was in fact the source of the problem: there was confounding between correlated factors leading to indeterminacy in the models. Therefore, we chose to model each factor in isolation and then compared the scaled coefficients (multiplying the mean value of the factor by its fit coefficient) to one another to gauge the relative influence each factor had on observing a non-reference allele, as well as, the influence each factor had on the size of the observed indel.

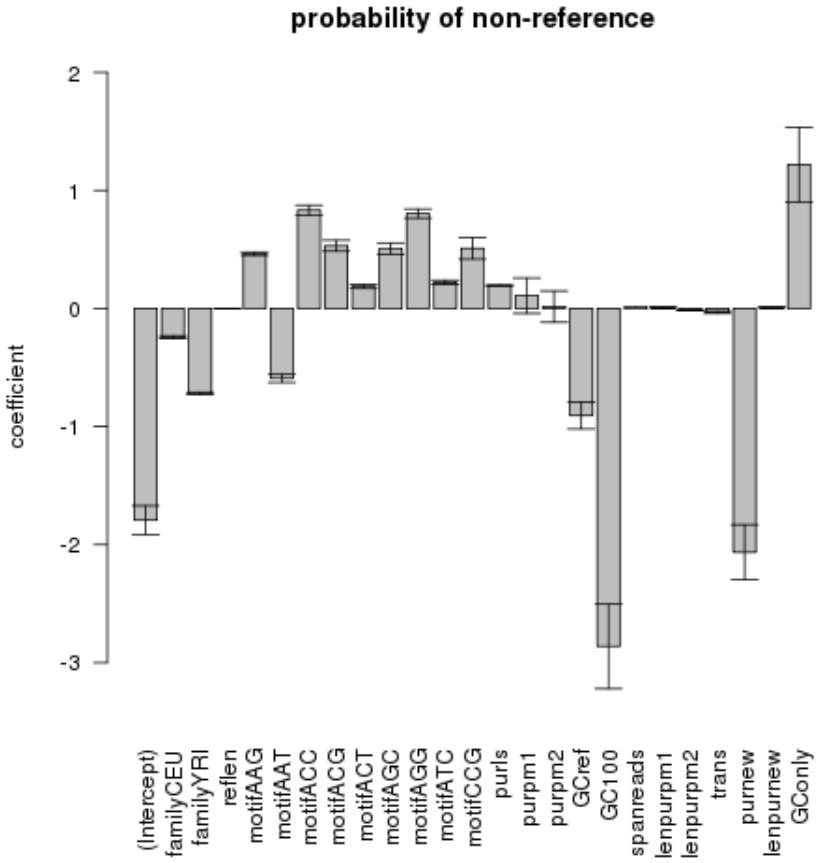


Figure 3.2: Graph of coefficients determined by full logistic regression of factors giving contradictory results because of confounding between correlated factors.

By sorting the scaled coefficients by the absolute value and plotting them on the same graph, it was clear which factors had the largest effect (be it positive or negative). In the end, we ended up with four plots: logistic regression for non-reference, linear regression for the magnitude of an indel and linear regressions for the size of both insertions and deletions. The graphs of each of these scenarios are plotted in figures 3.3, 3.4, 3.5 and 3.6 which illustrate the absolute effect of each of the factors. On each graph, all the coefficient values are shown aside from those having a p-value > 0.05 which include motifs ACG and AGC in the logistic linear model, motifs ACG and ACT in the insertions linear model and trans in the deletions linear model. Out of all the factors' coefficients that were graphed, all had a p-value < 0.001 except for GCref in the insertions linear model that

had a p-value in between the range of $(0.01, 0.05)$.

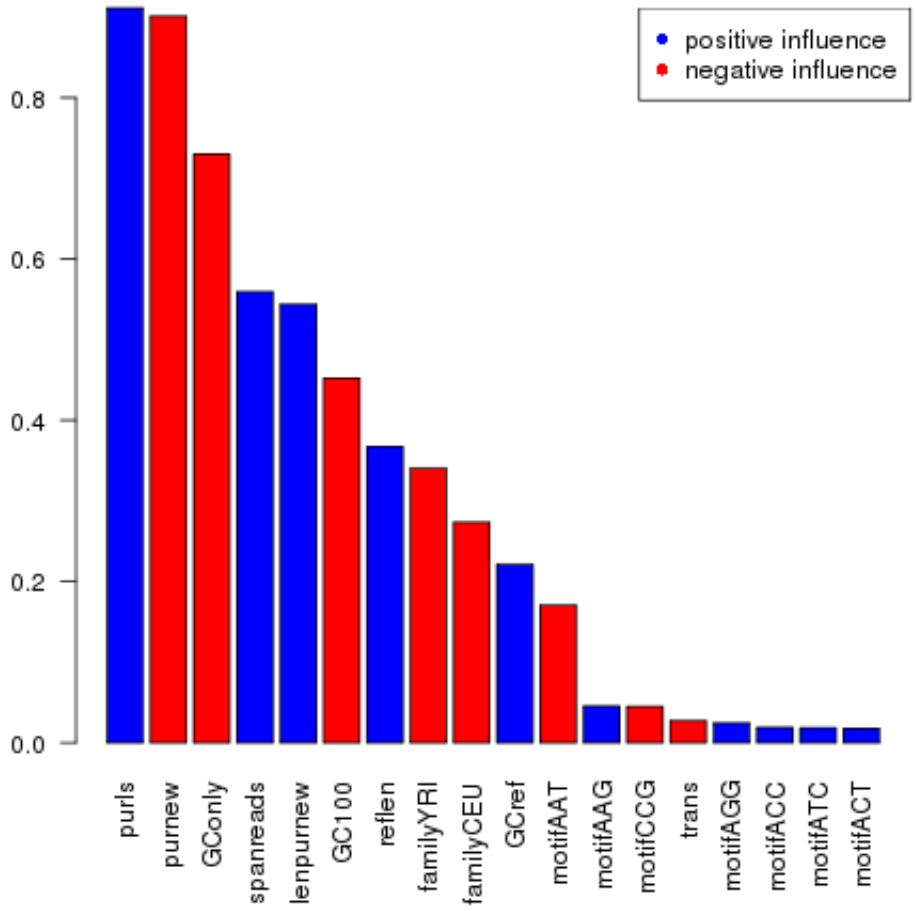


Figure 3.3: Bar graph of absolute values of coefficients from a logistic linear model for a STR being non-reference.

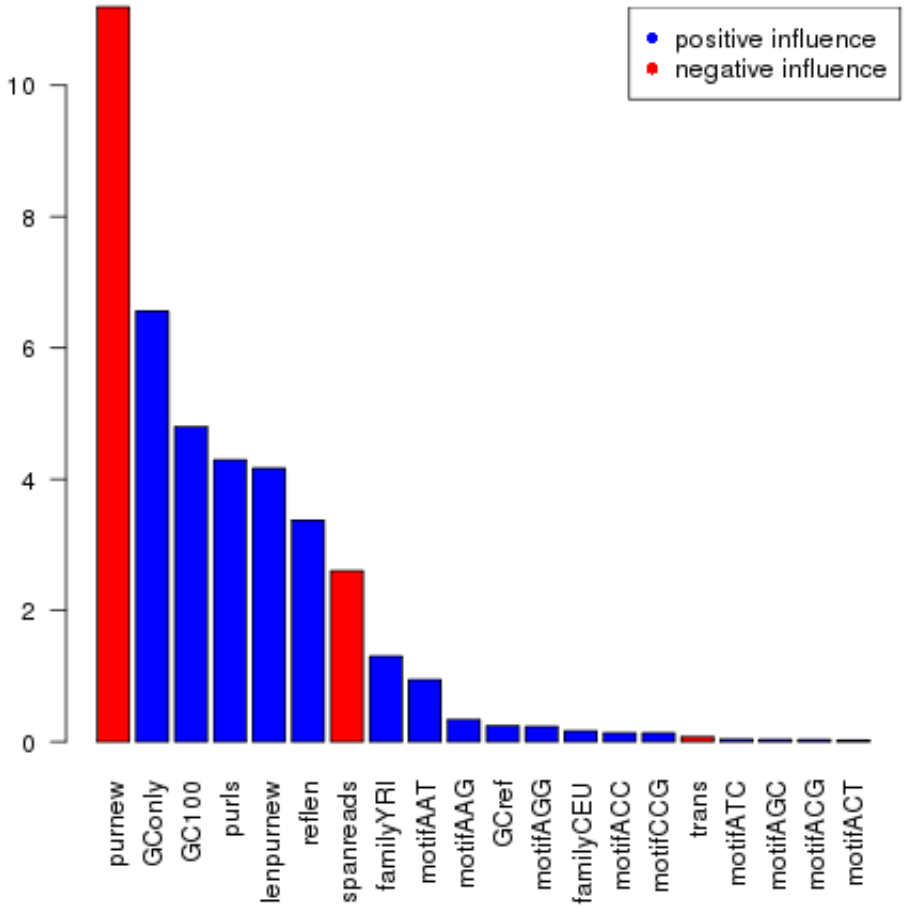


Figure 3.4: Bar graph of absolute values of coefficients from a linear model for the magnitude of an indel at variant STR loci.

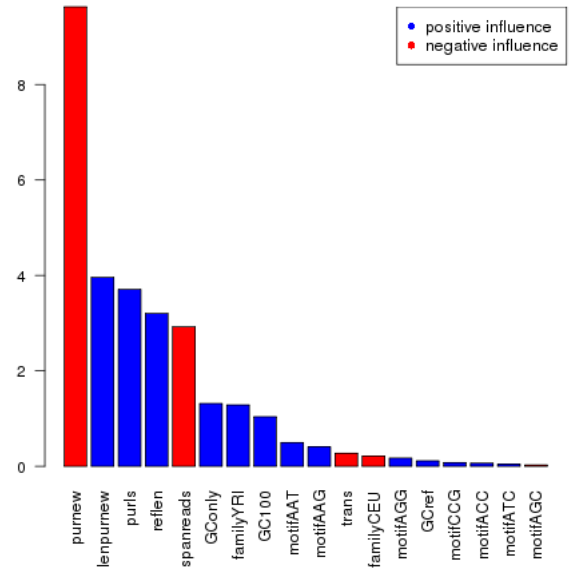


Figure 3.5: Bar graph of absolute values of coefficients from linear model for the magnitude of an insertion at variant STR loci.

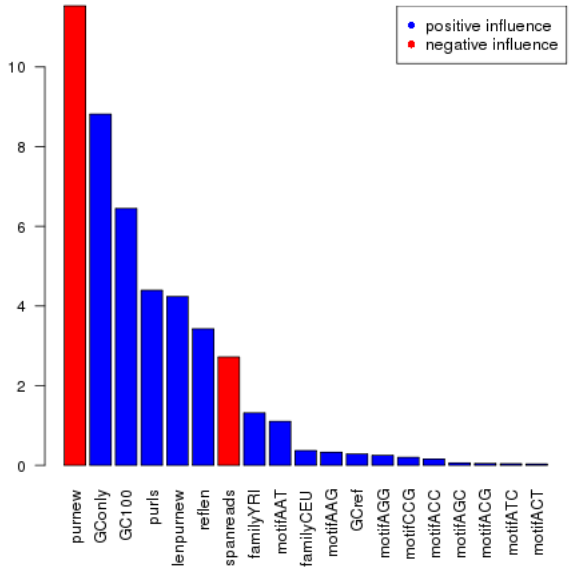


Figure 3.6: Bar graph of absolute values of coefficients from linear model for the magnitude of a deletion at variant STR loci.

3.5.1.1 Bias in modeling of purity

Upon inspecting the results of our regressions, it was surprising that the purity measure appears to be negatively correlated to the probability of observing a variant. Previous studies suggest that a higher purity increases the chance of mutation and polymorphism. Additionally, we found that the length of the longest pure subsequence in a repeat locus had the strongest correlation with observing a variant. We believe the cause of this correlation in opposition of what we would expect is that the the purity metric does not take into consideration the lengths of the STR in the reference. This would lead towards a bias of smaller repeats having a higher purity score than larger repeats because the chance of observing a foreign base or indel in a longer repeat is higher than a shorter repeat. Further, the criteria by which we ran our TRF means that shorter tandem repeats were not allowed to have any non-motif matching bases, otherwise they were not considered STRs. In order to test this belief, we graphed each locus's purity as a function of its length. Each STR locus was grouped into a bin of length 10 bp ranging from 15 to 205 bp. The values of these bins were then calculated showing a decrease of average purity as the repeat length increases. By simply multiplying the purity score by the repeat length in the reference, this bias is corrected.

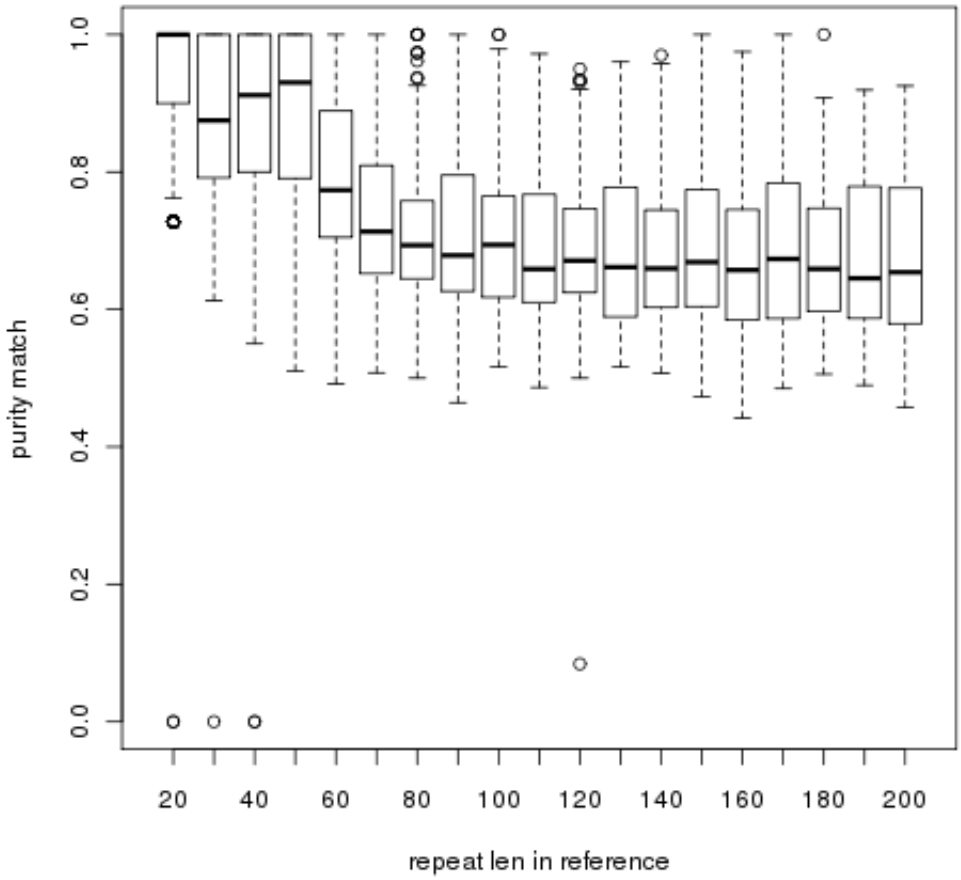


Figure 3.7: Boxplot of repeat purity across varying repeat lengths. This boxplot shows the values for our purity metric described in 3.4.4.2.1.

3.6 Discussion

Building upon our previous work in chapter 2, we have explored the effect a number of factors have on the probability of observing a variant in a STR locus. Using our previously described genotyping method for STRs, we ran a full genome analysis across nine deeply sequenced individuals – three trio data sets from two distinct populations (CEU, YRI) from the 1000 Genomes Pilot study and

Illumina's sequenced YRI trio. We made calls at 101,727 sites (sites having ≥ 10 spanning read pairs) across these nine individuals; 47,937 sites within this call set contained an observed variant in at least one of the alleles. Our method, as described in chapter 2, yields a very small number of false positives and when a variant is called, the variant's true length is almost always within a couple of repeat motifs' length of the actual repeat length in the resequenced sample. Because of this, we expect that any correlation we make is not coming from numerous spurious calls. The large number of calls across multiple loci ensures adequate power for our model, even if individual call sets are incomplete.

3.6.1 Sample family correlations

We decided to model some of the factors which might not be as interesting biologically, but that give us insight as to whether the actual correlations are correct, an *ad hoc* control so to speak. For instance, the family that a sample belongs to (CEU or YRI in 1000 Genomes Pilot Study, Illumina's trio) can increase or decrease the rate of observance of indels, because observing an indel is directly correlated with the sequence depth (see section 3.1.1.1) and overall shape (mean, standard deviation, skewness and kurtosis; see table 3.3) of the distribution. It is therefore not surprising that the Illumina trio has the most calls. This explains why the factors familyCEU and familyYRI have a strong negative influence in figure 3.3 (due to detection power) but much less and even an opposite effect in figure 3.4 which models the variant length conditional on the detection of a variant. This suggests that STRYPE's length estimates are not subject to read bias based on sequencing depth conditional on making a call.

3.6.2 GC composition correlations

Ignoring factors believed to be unimportant biologically or biased (family and length independent purity metrics), what was left were the true set of factors that play some sort of biological role in observing an indel at a given STR locus. Looking at figure 3.3, one of the largest influences on observing a variant is the amount of GC content proximal to the STR locus; the higher this GC content, the less likely you are to observe a variant. It is perhaps surprising that it is the GC

composition of the flanking regions rather than the repeat sequence itself that has one of the largest effects overall, as well as the largest amongst the GC content metrics. One technical explanation as to why fewer variants are observed in regions with high proximal GC content is the higher portion of sequencing errors in this region could lower the number of spanning read pairs, in turn lowing the power of our model to detect variants. In order to explore the external factor of mapping bias in the the genome, we compared directly the proximal GC content (GConly) to the GC content within the STR (GCref) based on the number of spanning reads observed at a given locus (see figure 3.8). The difference in the two metrics across the number of spanning read pairs showed that there is an indication that a higher GC content in the proximal sequence is associated with fewer spanning read pairs. For all but two bin sizes (190 and 200, which are the two smallest bins), the amount of flanking GC composition is anywhere from 10% to 35% higher than the STR composition, with the lower spanning read counts showing the strongest bias – which are also the largest bins.

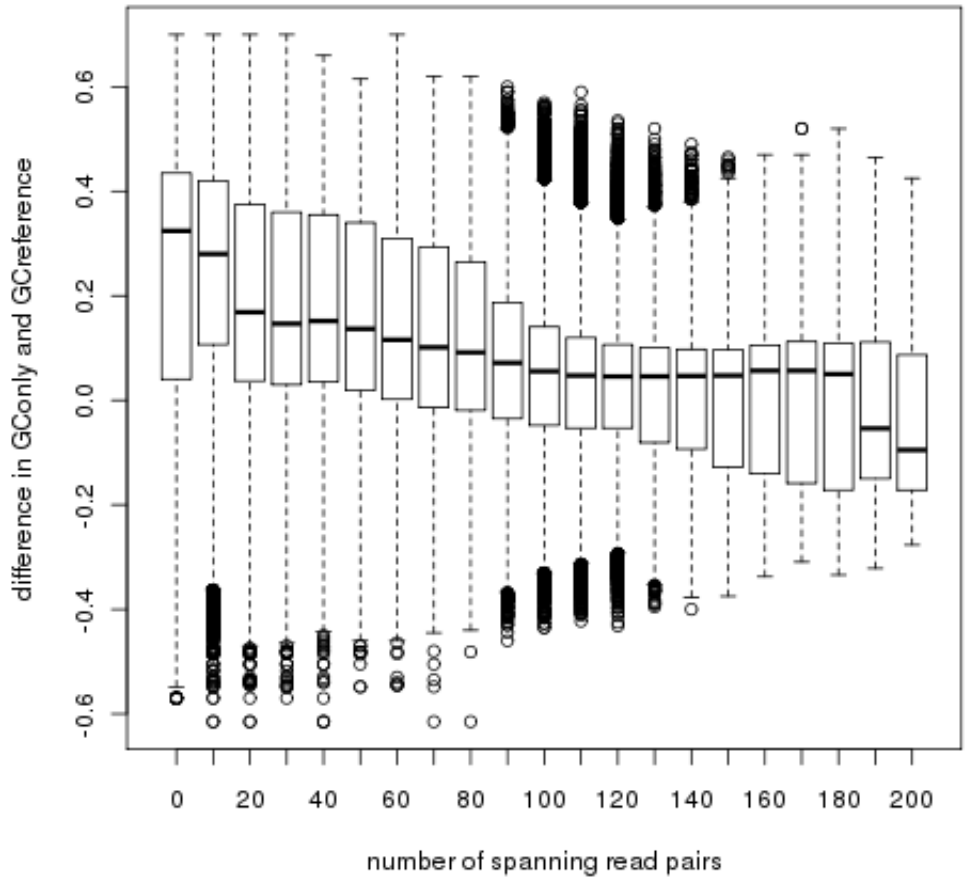


Figure 3.8: Boxplot of differences in GConly and GCref at a locus binned by the number of observed spanning read pairs at a locus. Each bin represents all sites in the genome which have a given number of spanning read pairs independent of the length.

3.6.3 Motif correlations

All but two motifs, CCG and AAT, were positively correlated with observing a variant (compared to the AAC family). While the families AAG, CCG, AGG, ACC, ATC and ACT all have comparable influence compared to one another, AAT has approximately five times more influence than the next strongest fam-

ily. As the family AAT is the only family to not contain any GC content, this correlation is in agreement with the factor GCref which is strongly correlated in the opposite direction. Most astonishing is while GC composition in the reference is positively correlated with observing an indel, the motif family CCG is negatively correlated. It might be possible that CCG repeats form some sort of secondary structure such as G-quadruplexes which are relatively prevalent in the genome and may decrease the chance of those sites undergoing mutation (Hazel et al. [2004], Huppert and Balasubramanian [2005], Bugaut and Balasubramanian [2008]). This is something to be explored further.

3.6.4 Purity correlations

The purity related correlation that had the largest effect out of all the factors was the length of the longest pure repeat in a locus. This correlation showed that the chance of observing a variant at a locus is less contingent upon the repeat's overall adherence to the motif than it is to the actual length of the longest pure stretch. Foreign bases and small indels which disrupt the motif frame may lower the rate of slippage, as well as other mechanisms that cause mutation at STRs discussed in chapter 1.

3.6.5 Further correlations: number of spanning read pairs, repeat length in reference and located within a transcript

The number of spanning read pairs, unsurprisingly, had one of the highest influences on whether a variant was observed at a repeat locus. Because of the design of our model, its clear that the more spanning read pairs at a locus, the more power there is to call a variant.

The length of a STR in the reference is also strongly correlated with observing an indel. This finding is in stride with the general understanding that longer stretches of DNA have a larger possibility of containing a variant. This correlation is directly in unison with the strongest indicator (purls) in that longer repeats in

the reference are also more likely to have longer pure stretches.

Lastly, there is a negative correlation of observing variants within transcripts. Our analysis in chapter 2 of indels called from capillary alignment showed that most triplet repeat variants were a multiple of three in length, which if occurring in an exon, would not disrupt the reading frame. However, the addition/removal of a multiple of three bases would in turn add or delete the number of multiples of three amino acids in a protein. Though not as detrimental as a reading frame shift, indels within a transcript (especially in an exon) are likely to be under purifying selection. Another possible contribution to the reduction of indels within transcripts is transcription-associated repair (Hanawalt [1994], Hoeijmakers et al. [2001]).

3.6.6 Independent analysis and comparison of each factors' effect on the magnitude of a variant at non-reference loci

A natural progression from the previous analysis is to determine the effect of factors on indel size at STR loci (see figures 3.4, 3.5 and 3.6).

3.6.6.1 All variants

Many of the correlations seen in the logistic linear model are the same as in the linear model for indel magnitudes. If a factor is positively correlated with observing a variant, it is also positively correlated with the size of the variant. However, the proximal GC content (GConly, GC100) is now strongly correlated with observing larger indels while it is negatively correlated with observing a non-reference locus. This can be explained by the lower amount of spanning reads when proximal GC content is high (see 3.6.2). Smaller size variants would need more spanning reads to be called while larger variants need less. Therefore, the larger variants would be more readily called in regions of high GC content.

The entirety of motif families are also positively correlated. AAT has the largest effect for observing a larger indel but from the previous analysis, is negatively

correlated with observing a variant (the largest coefficient of the motif families). This is quite interesting. Motif CCG also exhibits this interesting reversal.

As expected from our modeling of non-reference variants, the length of the longest pure stretch and reference positively effects the magnitude of the variant when one is observed. Larger indels are more likely in longer STRs because there is more sequence which can undergo replication slippage compared to shorter STRs.

The number of spanning reads also exhibits a reversal in influence from observing a variant to the size of the variant. While observing a variant is strongly influenced by the number of spanning reads, the number of spanning reads actually decreased the magnitude. As longer variants reside in longer repeats, these loci inherently have less spanning reads. Additionally, as discussed earlier, larger variants need less spanning reads to be called as the signal is stronger than smaller variants. This would explain why the number of spanning reads is negatively correlated with observing a variant.

Lastly, residing within a transcript is negatively associated with observing larger indels. The larger the variant within a transcript, the more disruptive it will be, especially if it resides in the exon which will affect the production of the amino acid chain during translation.

3.6.6.2 Independent analysis of insertions and deletions compared to the reference

When comparing the magnitude of indel calls in insertions versus deletions, almost all correlational directions match one another with the exception of the motif family AGC which is negatively correlated in inserts and positively correlated in deletions. Its effect, however, is relatively small in both directions and is most likely statistically insignificant. The correlations that stand out the most are in the same relative order of significance. While the strongest indicators of larger variants are the purity metrics for insertions, it is the proximal GC content for deletions. All other factors seem to be in the same order and relative influence to one another. A simple explanation is not readily available and warrants

further analysis. It should also be noted that the reference genome does not represent the ancestral state. Many of the tandem repeats were estimated using BACs and so at variable loci the allele present in the BAC was chosen, which typically will represent a selection at random according to the population allele frequencies. This makes inference difficult when comparing whether insertions or deletions are more likely as we can not say for sure that the alleles in the reference represent the ancestral state.

3.7 Conclusion

We have seen evidence for a variety of effects on STR mutation properties that are broadly in line with previous expectations (Kelkar et al. [2008]). Aside from the independent correlation values, the knowledge of which factors have the strongest effect could assist in our future modeling of STR indels. We could use this information to describe a more accurate prior than the one we developed in chapter 2.

Chapter 4

Population based analysis of short tandem repeats

Collaboration note *This chapter contains work performed in collaboration with David Knowles. David assisted in developing the statistical machinery used in estimating the allele vector at each STR locus.*

As sequence depth plays the most important part in our ability to assay variation in STRs using short paired end reads, STYRPE is restricted to genotyping only individuals who have been sequenced to a relatively high physical coverage depth. However, a major mode of current genome wide sequencing is to sequence many individuals from a population at a lower depth – as in the 1000 Genomes Project (Consortium [2010]) and the UK10K (www.uk10k.org). For example, the target 4x depth that the 1000 Genomes Project is using for genome wide sequencing is well below what is necessary for our model to make informative calls on a single individual's genotype at a STR.

However, within the spectrum of population genetics, each locus in a diploid individual is comprised of two alleles which are more than likely shared across numerous individuals in that population. If we could use the combined information from multiple individuals, we would have enough sequence information to make predictions of the overall frequency of alleles at a locus, as well as how diverse a locus is. This would complement our analysis of factors which affect

the chance of observing an indel at a given STR, as well as give us a list of candidate sites which might be multiallelic (characterized by many alleles) or whose underlying allele frequency in a population is not best described by the reference allele length. What this essentially means is: does some number of individuals in a population have the reference allele, or is/are there an alternate set of allele(s) at that locus comprising a certain density not coinciding with the reference allele length.

4.1 Low coverage individuals in the 1000 Genomes Project

As briefly described in chapter 1, the 1000 Genomes Project is a massive, multinational sequencing project which endeavors to sequence 2,500 individuals across twenty-seven populations. In the intermediate data sets that we consider here, corresponding to an early phase I freeze from November 2010, 929 individuals were sequenced with the Illumina paired end read platform that had at least one library that passed quality control requirements. In all, 1,122 libraries have been sequenced which pass the quality control criteria (about 1.2 libraries per individual).

4.1.1 Sources of sequence

Sequenced at multiple centres, each individual's sequence was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/ having been mapped to the human reference genome, GRCh build 37, using the BWA alignment tool.

4.1.2 Sequencing statistics

The sequencing coverage was calculated for every library/individual (as in previous chapters) for those sequenced in the 1000 Genomes Project. Each library was sequenced to a much lower depth than in the previous chapters, ranging from a library sequencing coverage of 0.010 to 10.299x (mean of 2.456±1.456) and an individual sequencing coverage from 0.0096 to 10.756x (mean of 2.966±1.464). This

is lower than the target coverage for the project of 4x per sample because this is an interim data set and we included all samples with any sequence, however little.

In total, fourteen populations were sequenced ranging in number of individuals from 6 to 98, as well as, the number of libraries per population ranging from 6 to 122. Thirteen of the fourteen populations had a combined depth greater than 170x, with the deepest coverage coming from the JPT population at 303x. The largest population, TSI, had an amalgamated base coverage of 235.961x. This would mean, given an allele of frequency 20% in a population, you would have an effective depth of 38.1x which should be sufficient to detect it (depth taken from median population sequencing depth of 190.489x). The power to discern set variants only increases as the number of individuals sequenced increases, contingent upon the samples having a shared allele amongst them. The statistics for each of the populations is listed in table 4.1.

Population statistics for 1000 Genomes Project low coverage data set						
Population	Individuals	Libraries	Bases sequenced	Coverage	Avg. cov. (lib)	Avg. cov. (ind)
YRI	66	74	628904103390	209.635	2.833	3.176
ASW	50	57	544728006968	181.576	3.186	3.632
GBR	70	90	519154431151	173.051	1.923	2.472
TSI	98	122	707881625221	235.961	1.934	2.408
CHB	81	141	582563154053	194.188	1.377	2.397
CLM	50	50	518391563974	172.797	3.456	3.456
LWK	83	93	783360704228	261.120	2.808	3.146
MXL	54	59	540194155266	180.065	3.052	3.335
CHS	92	104	672006329021	224.002	2.154	2.435
PUR	52	59	560368488942	186.789	3.166	3.592
JPT	72	77	911055671783	303.685	3.944	4.218
IBS	6	6	49644449600	16.548	2.758	2.758
FIN	75	90	548432746738	182.811	2.031	2.437
CEU	80	100	700341495829	233.447	2.334	2.918
totals	929	1122	8267026926164	2755.675	2.456	2.966

Table 4.1: Summary of sequencing statistics for individuals' libraries in 1000 Genomes Project low coverage data set. The number of individuals per population ranges from 6 to 98 and number of libraries per population ranges from 6 to 122. The total number of bases sequenced from each individual/library is summarised in the fourth column with the average per base coverage across all individuals in the fifth column. The last two columns indicates the average base coverage per library and individual in each population, respectively.

4.1.3 Population MPERS distributions

A major component of our analysis is based on the concept that each individual belongs to a local population and that their alleles will be drawn from an unobserved distribution of alleles from within these populations. This means that in principle: the more individuals there are in a population sample, the more power there should be to detect the underlying allele frequencies and general dispersion of STR lengths within a loci. In a global population analysis, however, the alleles might be so dispersed that it becomes hard to resolve one from another. Before we carried out any further modeling, it was important to look at the distributions of MPERS across the 1000 Genomes libraries to get an estimate of the general distributions of fragment sizes.

Given that there are over a thousand libraries sequenced for the 1000 Genomes Project, there is not much we can really deduce from the plot of all MPERS distributions (figure 4.1). However, libraries which differ in fragment length but maintain similar variances are almost identical in terms of information they are are able to give. Larger libraries will be able to assay longer STRs and at equal coverages, yield more spanning read pairs, but for a STR of length less than both fragment libraries, each library will supply approximately the same amount of information per spanning read pair. We therefore centered these distributions by offsetting their mean to zero and compared the more important characteristics of the distributions such as variance and shape (figure 4.2). The general form of a unimodal distribution across all libraries is promising in the context of genotyping STRs across populations (figure 4.2). Again, the sheer number of distributions does obfuscate the assessment of the general shape of each library's distribution as the magnitude of overlying distributions is not explicitly shown. When we break the libraries down by population, it becomes clear which populations are more informative in terms of shape and distribution.

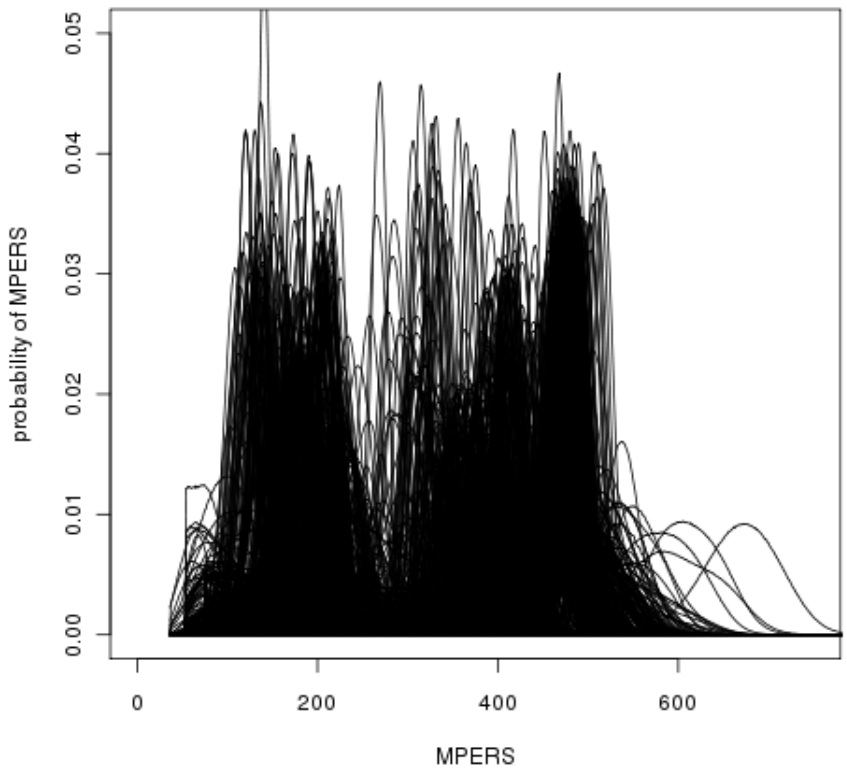


Figure 4.1: Plot of MPERS distributions for every library in the 1000 Genomes Project data set. Most libraries' MPERS fall within the range of 150 to 600 bp with peaks (fragment library sizes) around 150, 200, 400 and 500 bp.

4.2 Modeling

Modeling a population's underlying distribution of alleles within a STR locus relative to the reference adds multiple complexities compared to the previous modeling of a single individual's genotype as described in chapter 2. Instead of assuming that all spanning reads come from a maximum of two alleles, now the union of all indels in the population is possible.

We still take a Bayesian approach, calculating the (log) likelihood of the observed data, and combining a prior with this to estimate the posterior. The

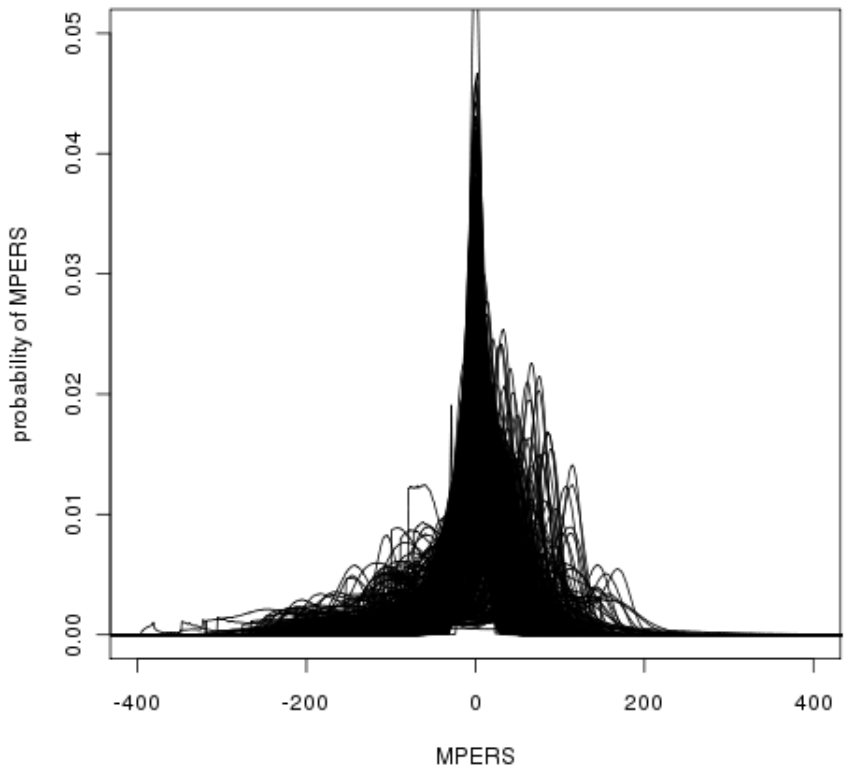
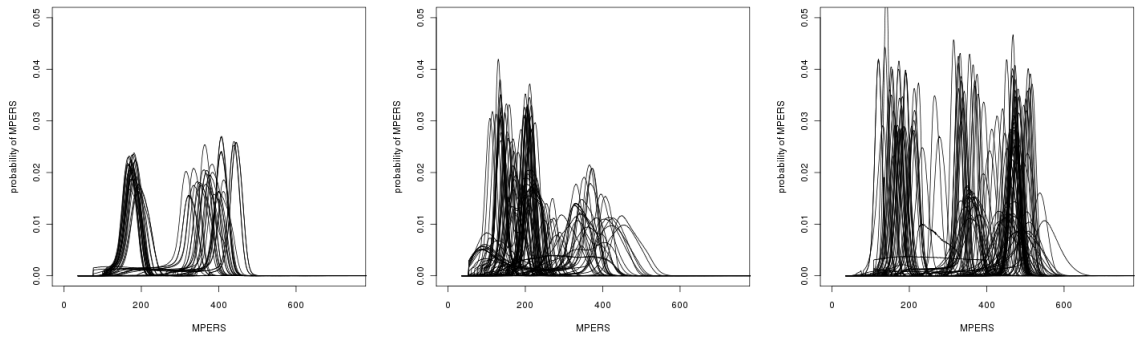


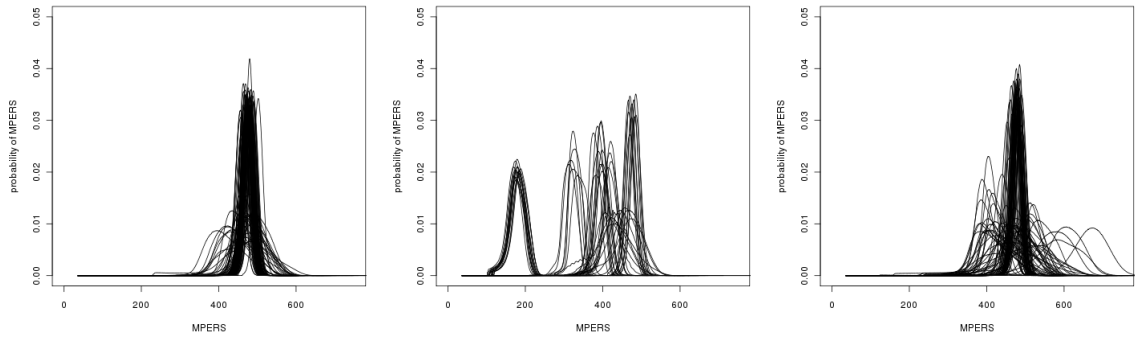
Figure 4.2: Plot of MPERS distributions whose mean of each library is arbitrarily set at zero. It is clear that the majority of libraries have a MPERS variance that is tightly bound around the mean value (peak at zero). This does not mean that all libraries behave well (as signified by the MPERS distributions whose values fluctuate highly away from the mean). However, the prevailing shape of the MPERS distributions tend towards an adherence to being tightly bound.

matrix likelihood is calculated from the full likelihood matrix from each individual across the set of possible diploid calls at a site([-30x30],[-30,30] in the case of triplet repeats) exactly as in chapter 2.

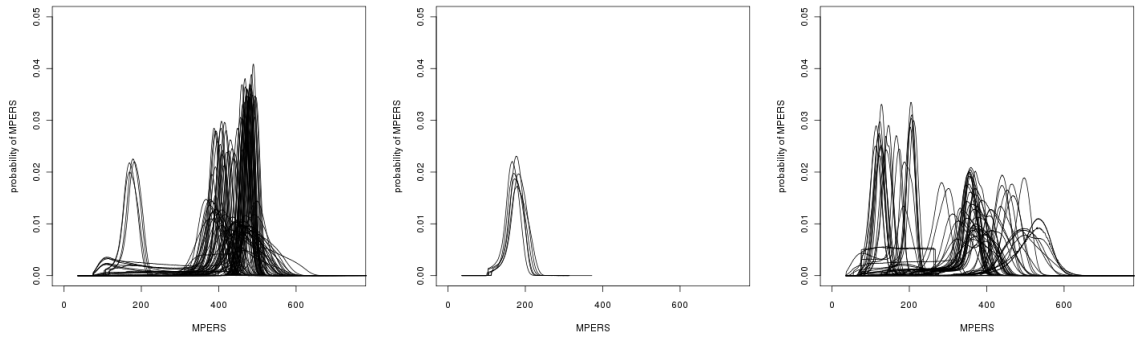
Not interested in a specific individual's genotype, our previous prior over diploid genotypes from chapter 2 was no longer appropriate. Instead, we needed a prior over all distributions of alleles at a locus. As we were no longer looking for genotypes, but allele frequencies, it meant that our posterior would take the form of a



(a) MPERS for ASW population (b) MPERS for CEU population (c) MPERS for CHB population

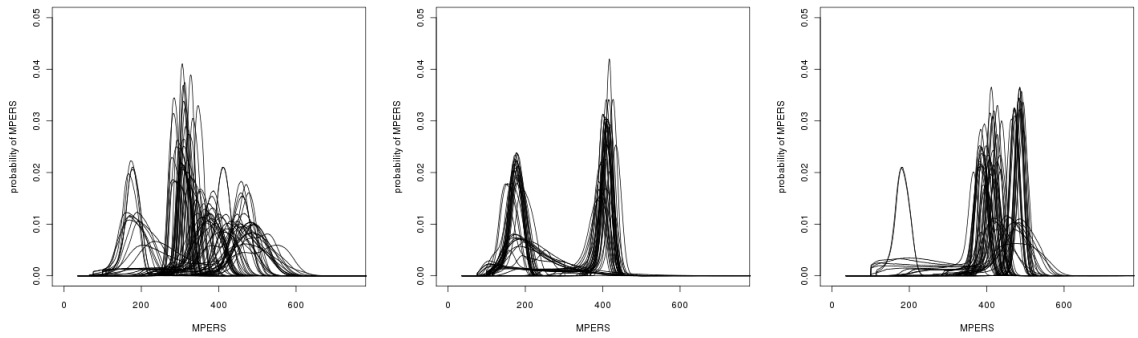


(d) MPERS for CHS population (e) MPERS for CLM population (f) MPERS for FIN population

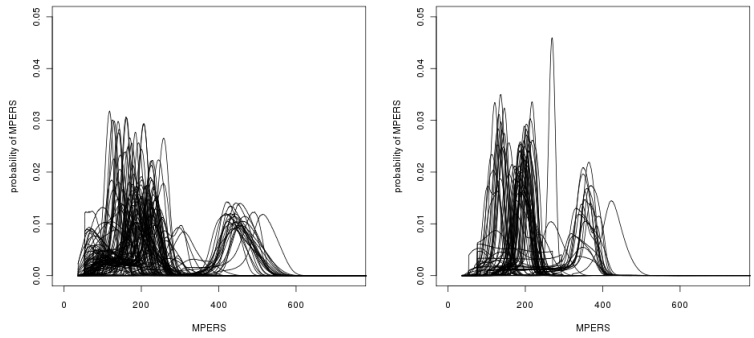


(g) MPERS for GBR population (h) MPERS for IBS population (i) MPERS for JPT population

multinomial distribution; where each indel value in the multinomial distribution was representative of the relative frequency of that allele within the population. Achieving a multinomial posterior distribution meant that we would use a Dirich-



(j) MPERS for LWK population (k) MPERS for MXL population (l) MPERS for PUR population



(m) MPERS for TSI population (n) MPERS for YRI population

Figure 4.3: Plots of the raw MPERS for each of the fourteen populations in the 1000 Genomes Project data set. Each population is usually sequenced by libraries having multiple fragment size libraries (with an exception of CHS, FIN and IBS).

let prior. The Dirichlet distribution is the conjugate prior for the multinomial distribution and is made up of a family of continuous multivariate probability distributions parameterized be a single vector α. The Dirichlet probability density function returns the belief that the probabilities of $|K|$ mutually exclusive events are x_i given that each event has been observed $\alpha_i - 1$ times. The values of vector α represent the number of pseudo counts for a given event x_i.

The Dirichlet distribution of order $|K| \geq 2$ having parameters of $\alpha_1, ..., \alpha_{|K|} > 0$

has a probability density function given by

$$f(x_1, \ldots, x_{|K|-1}; \alpha_1, \ldots, \alpha_{|K|}) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{|K|} x_i^{\alpha_i - 1} \tag{4.1}$$

for all probabilities of vector X $(x_1, ..., x_{|K|})$ being non-zero, positive and satisfing the condition that $x_1 + ... + x_{|K|-1} < 1$, where x_K is simply the probability calculated directly as $1 - x_1 - ... - x_{|K|-1}$ and the density is zero outside this open $K-1$-dimensional simplex. The distribution is normalized by the multinomial β function.

Because we normalize to obtain posteriors, in practice we could drop the β function and use a proportional Dirichlet prior as the values will correlate directly to the actual probabilities described in equation 4.1. The Dirichlet prior's parameter vector α will consist of $|K|$ possible indel values. The probability of any one of these values is p_k. The vector $\mathbf{p}$ is a probability vector whose elements are all > 0 and sum to one. Therefore, our Dirichlet prior is expressed as

$$\pi(\mathbf{p}) \propto \prod_{i=1}^{|K|} p_i^{\alpha_i - 1}$$

Looking now at population alleles instead of genotypes, we will assume within a population – and by extension an individual (n) – all indels (i) are mutually independent of one another such that

$$p(I_1, I_2) = p(I_1) \cdot p(I_2) = p_{I_1} \cdot p_{I_2}$$

Next we define the conditional distribution for a purported population allele vector $(\mathbf{p})$ for genotype calls in an individual as

$$p(I_1, I_2, d_n|\mathbf{p}) \propto p(d_n|I_1, I_2) \cdot p(I_1, I_2|\mathbf{p}) \tag{4.2}$$

$$p(d_n|I_1, I_2) = l_{n,I_1,I_2}$$

$$p(I_1, I_2|\mathbf{p}) = p_{I_1} \cdot p_{I_2}$$

where d_n is all the spanning read information for an individual at a given locus and l_{n,I_1,I_2} is the likelihood of the data in individual n having genotype $\{I_1, I_2\}$ as calculated in chapter 2.

Having defined the joint probability distribution for an individual, it was not obvious the best means by which we should model this system. We sought methods capable of learning the best values of $\mathbf{p}$ from the data, which essentially represents the true underlying frequency of alleles at a locus in a population. In the end, we choose two different algorithms to explore; the Expectation-Maximization algorithm (EM algorithm) and Gibbs sampling. However, first we will describe the priors that we used.

4.2.1 Priors

We considered three priors ($\pi()$) for our modeling which had the following initialization parameters α (pseudcounts)

1. **uniform**: a uniform prior with an α value of one for every indel size
2. **conservative**: a prior with 0.8 of the weight on the reference allele (α of 80) and the rest of the weight equally distributed across the indels, 0.01 (α of 1.
3. **decay**: a prior used in chapter 2 where the most weight is on the reference allele and then a gradual decay of weight as indel sizes move away from the reference, pseudo counts found by multiplying the probability of an indel by 100

4.2.2 EM algorithm

The EM algorithm is a method for determining either the maximum likelihood or maximum *a posteriori* (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables – which in our case

are the underlying frequencies of indel alleles in a population. The algorithm takes an iterative approach which switches between performing the expectation step (E) and the maximization step (M). In the E step, the algorithm computes the expectation of the log-likelihood evaluated with the current estimate for the parameters (the indel allele frequency in the population), then the M step recalculates the parameters which maximize the expected log-likelihood found in the E step. The new parameter values found in the M step are then used in the next iteration of the E step and this process is iterated, hopefully converging at the true parameter values.

The problem with MAP inference is that it ignores the uncertainty in our indel assignments. For the high coverage samples in chapter 2, this is not as much of a problem as we were solely interested in the genotype of a single individual and had enough power to make a genotype call. But for the low coverage individual's in the population, this is more of a problem as we may be over fitting the data. Instead we keep the posterior distribution for each individual; $q_n(i_1, i_2)$.

For purposes of inference, it is convenient to write the prior on the allele frequencies as

$$p(i) = \prod_K p_k^{\mathbb{I}[i=k]} \tag{4.3}$$

where $\mathbb{I}[i = k]$ is the indicator function; this interpretation was used for ease of computation in the EM model shown later. Similarly, it was convenient to write the likelihood terms for individual n in the same form

$$L_n(i_1, i_2|d_n) = \prod_{s \in I_1, t \in I_2} l_{n,s,t}^{\mathbb{I}[i_1=s, i_2=t]} \tag{4.4}$$

When these two equations (4.3 and 4.4) are combined with equation 4.2, the joint distribution for the population model is

$$p(\mathbf{p}, \{i_{1,1}, i_{1,2}, ..., i_{|N|,1}, i_{|N|,2}\}, D) \propto \prod_K p^{a_k - 1} \cdot \prod_N \prod_{s \in I_1, t \in I_2} [p_s^{\mathbb{I}[i_1=s]} \cdot p_t^{\mathbb{I}[i_2=t]} \cdot l_n^{\mathbb{I}[i_1=s, i_2=t]}] \tag{4.5}$$

which in log space would be

$$\log p(\mathbf{p}, \{i_{1,1}, i_{1,2}, ..., i_{|N|,1}, i_{|N|,2}\}, D) \propto \sum_K (a_k - 1) \log p_k + \sum_N \sum_{s \in I_1, t \in I_2} \mathbb{I}[i_1 = s] \log p_s + \mathbb{I}[i_2 = s] \log p_t + \mathbb{I}[i_1 = s, i_2 = t] \log l_{n,s,t}$$

As it is apparent here, having written the terms in the form of indicator functions, it is simple to take the expectation with respect to q to give the following variational lower bound on the log marginal likelihood as

$$l(q, p) \propto \sum_K (a_k - 1) \log p_k + \sum_N \sum_{s \in I_1, t \in I_2} q(i_1 = s) \log p_s + q(i_2 = t) \log p_t + q(i_1 = s, i_2 = t) \log l_{n,s,t}$$

where $q(i_1 = s) = q(i_2 = s) = \sum_t q(i_1 = s, i_2 = t)$ which aggregates all the mass of the two-dimensional matrix (genotype calls) into a one-dimensional vector representing the overall frequency of an allele in a population at a given locus. The E step is now simply

$$q(i_1 = s, i_2 = t) \propto p_s \cdot p_t \cdot l_{n,s,t} \quad \forall s, t$$

It should be noted that q must be normalised such that $\sum_{s \in I_1, t \in I_2} q(i_1 = s, i_2 = t) = 1$. Finally, the M step will maximise parameters with respect to the prior as

$$p_k \propto \alpha_k - 1 + \sum_N [q(i_{n,1} = k) + q(i_{n,2} = k)] = \alpha_k - 1 + 2 \sum_N q(i_{n,1} = k)$$

where the final equation expresses the symmetry between i_1 and i_2.

4.2.3 Gibbs sampling

As a second, non-deterministic method, it would be useful to check the results of our EM algorithm by having the full posterior using a Monte Carlo Markov chain approach (MCMC). We used the Gibbs sampler for our MCMC process. In essence, the Gibbs sampler samples from the two latent variables $\mathbf{p}$ and I in hopes

of describing the true posterior. To start, we initialize $\mathbf{p}$ with some reasonable value (i.e. uniform, gradual decay in density as you move away from the reference length and equal dispersion of densities across the indels with the majority of the density on the reference). From the conditional distribution in equation 4.2, we can derive the conditional distribution for $i_{n,1}, i_{n,2}$ as

$$p(I_{n,1}, I_{n,2}|\mathbf{p}, d_n) = p_{n,i_1} \cdot p_{n,i_2} \cdot l_{n,i_1,i_2} \tag{4.6}$$

The sampling of $(I_{n,1}, I_{n,2})$ involves sampling from the two-dimentional discrete distribution for each individual n. Given the genotype $\{i_{n,1}, i_{n,2}\}$, the conditional distribution on $\mathbf{p}$ is a Dirichlet and is calculated as

$$\begin{aligned} p(\mathbf{p}|\{i_{1,1}, i_{1,2}, ..., i_{|N|,1}, i_{|N|,2}\}, D) &\propto \prod_K p_k^{\alpha_k - 1} \cdot \prod_{n \in N} p_{n,i_1} \cdot p_{n,i_2} \\ &\propto \prod_K p_k^{\sum_N (\mathbb{I}[i_1 = k] + \mathbb{I}[i_2 = k] + \alpha_k - 1)} \end{aligned} \tag{4.7}$$

which is another Dirichlet with parameters given by the summation in the exponent $(\sum_N (\mathbb{I}[i_1 = k] + \mathbb{I}[i_2 = k] + \alpha_k - 1))$. Explicitly, this equation is summing the number of allele calls of a particular allele size swithin a population at a given locus and combining these with the prior pseudocounts. We let the Gibbs sampler run which iterates back and forth between sampling from $\mathbf{p}$ and I using equations 4.6 and 4.7, respectively. We store each iteration's values which are later used to estimate our model's parameters.

4.3 Simulation

To compare the EM and Gibbs sampling approaches, we simulated data with various distributions of indel alleles, using real STR loci as our template. We selected these sites from the 1,881 triplet repeat loci found by TRF on chromosome 20.

4.3.1 Simulation of MPERS for spanning read pairs

The number of simulated spanning read pairs at each locus should match the number of spanning read pairs observed at the same locus in the real data. This ensures that our simulations will not give better results because of a discrepancy in the number of of spanning read pairs. Looking across all positions in chromosome 20 (1,881 loci), we determined how many spanning read pairs were at each locus for each individual's library as we had in chapter 2. The count of spanning read pairs was used to determine how many spanning read pairs we would simulate for each individual's library.

The separation sizes of spanning read pairs that we simulated depended on the empirical MPERS distribution of the relevant library, and on the repeat length of each locus. Each sequence library's length distributions were calculated from approximately ten million reads (as discussed in chapter 2), but as this set of MPERS does not adhere to the bias of MPERS in longer STRs, we sampled directly from the generated empirical distributions (see chapter 2). For example, if we were interested in simulating a scenario where all the individuals in a population contain the reference allele at both copies – say a length of 50 bp – then for each individual's library, we would sample some number of reads (as taken from the number of observed spanning read pairs in the real data) from distributions of length 50 bp. The distributions were comprised of the MPERS and the probability of observing that MPERS in the genome conditioned on the reads being drawn from a repeat length of length l. We sampled directly from this distribution by first calculating the cumulative distribution of the MPERS in rank of smallest to largest, and then randomly sampled a value between [0,1] with a precision of 10^{-7}, or the probability of sampling a single MPERS from the distribution. This value correlated within some range of the cumulative distribution of the MPERS (described as a step function) and the MPERS whose cumulative probability value was the closest was the sampled MPERS. We did this for each set of spanning read pairs for each individual's library. These MPERS were then used to calculate the likelihood of genotype calls for each individual as described previously in section 4.2.

The process becomes a bit trickier when we move away from simulating a homozygous reference scenario. First, we need to correctly simulate the relative frequency of an allele within a population. A simple example would be where fifty percent of all alleles in a population coincide with a deletion of 12 bp relative to the 50 bp reference length and the other fifty percent coincide with the reference allele. This means that each individual has a fifty percent chance that each of her alleles are either the deletion allele or the reference allele. This means that there are three possibly genotypes an individual can have: homozygous reference, homozygous indel and heterozygous. To simulate this, each individual is sampled twice from the frequency distribution of alleles at a locus. This yields the true genotype of the individual at that locus. Then for each spanning read pair (numbering in the amount of spanning read pairs in the real data as before), the allele from which the spanning read pair comes from is sampled at a fifty percent probability that it comes from either one allele or the other. This will obviously only have any meaning for individuals whose simulated genotype is heterozygous but it is important as the sampling of reads in real data is drawn at random from one allele or the other. This procedure is carried out for every individual's library such that each person has some count of reads being drawn from one of the two alleles that were sampled from the overall distribution of alleles in the population. The spanning read pairs are then sampled from the distributions of MPERS from an individual's library in the same form as described above but with one additional criteria: that the distribution from which the MPERS is sampled from coincides with the true STR length. For example, say an individual was sequenced from a single library and at a specific locus had four spanning paired end reads. From the sampling of alleles, it came out that this individual was heterozygous at this particular locus and it worked out that two reads came from the reference allele and two reads came from the deletion allele. This means that two MPERS were sampled from the distribution for that individual's library which coincided with the reference allele length (50 bp) and two MPERS were sampled from the distribution for that individual's library which coincided with the deletion allele length (50 - 12 bp or 38 bp). All four reads were then used in calculating the likelihood of genotype calls for that individual's library, where

the calculation of the likelihood is naive as to which allele these sampled MPERS were drawn from – as would be the case for real data.

4.3.2 Simulation results

The simulated reads were used as input into our two algorithms for three different scenarios: only reference alleles, two alleles off reference (±-9 bp, both at a frequency of 0.5) and three alleles (0.45 density on both alleles -12 bp and 6 bp and 0.1 density on the reference allele). We decided to look at multiple frequency distributions to be sure that our algorithms were able to work on all frequency scenarios we would encounter in real data. We also chose to use multiple populations to check the robustness of our model and to be sure that a model's efficacy is not contingent upon some unobserved criterion specific to a population. For our analysis, we decided to use populations CHS and CLM which are comprised of 92 and 53 individuals, respectively. Our simulations were conducted using a uniform prior which was a reasonable choice for our simulations to check whether each of the algorithms was overfitting the data or not. The uniform prior would not be appropriate for our later analysis of real data when we looked at the entropy, off reference and off ±3 bp for each locus in a population (discussed in section 4.4).

4.3.2.1 Reference allele frequency

The first simulation was on the CHS population from an allele frequency distribution that was entirely comprised of reference allele lengths. We randomly chose 14 loci in chromosome 20 for our analysis. We forced each locus's length in the simulation to match the reference and sampled MPERS from the distribution which coincided with the reference length. The vector values for these 14 sites for both the EM and Gibbs algorithm are shown in figures 4.4 and 4.5.

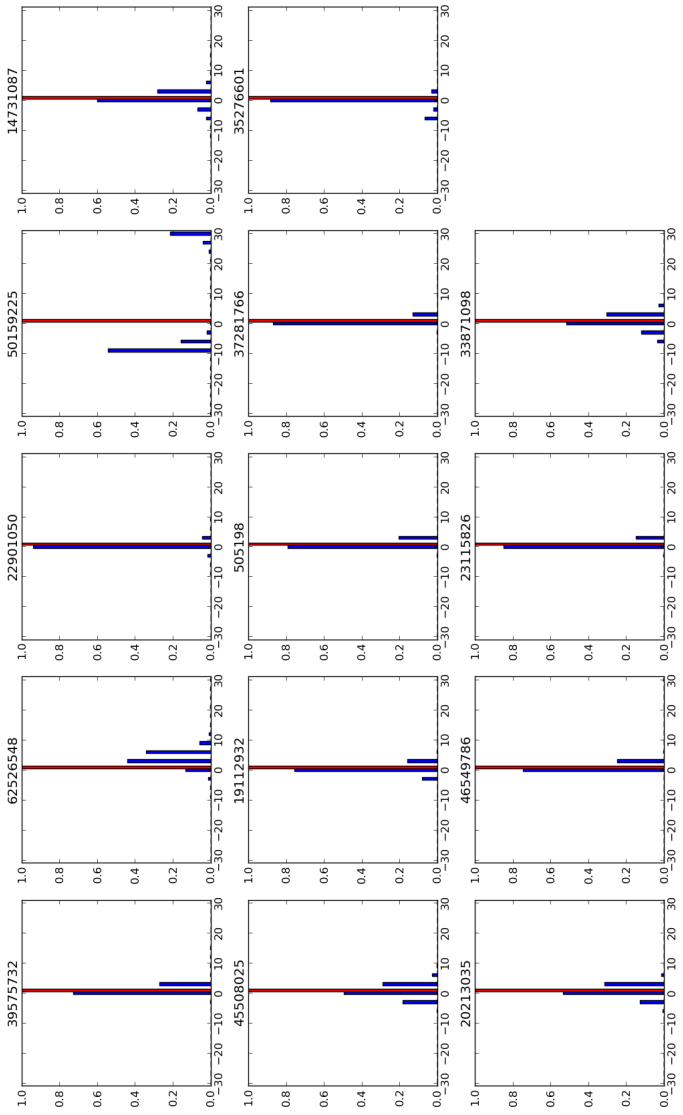


Figure 4.4: Prediction of allele frequency distribution for the EM algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely of reference alleles based on a CHS population (red bars). Most all the predictions' allele frequency distributions center around the truth (reference). However at start position 50159225, the predicted frequency allele distribution differs greatly from the truth. Further inspection showed that for this site, there were fewer reads spanning at this locus from the real data, which in turn meant fewer simulated spanning read pairs which the EM algorithm could use. Another example of misfitting is at position 62526548. In these cases, the EM algorithm can over fit the data, leading to a confident false positive call.

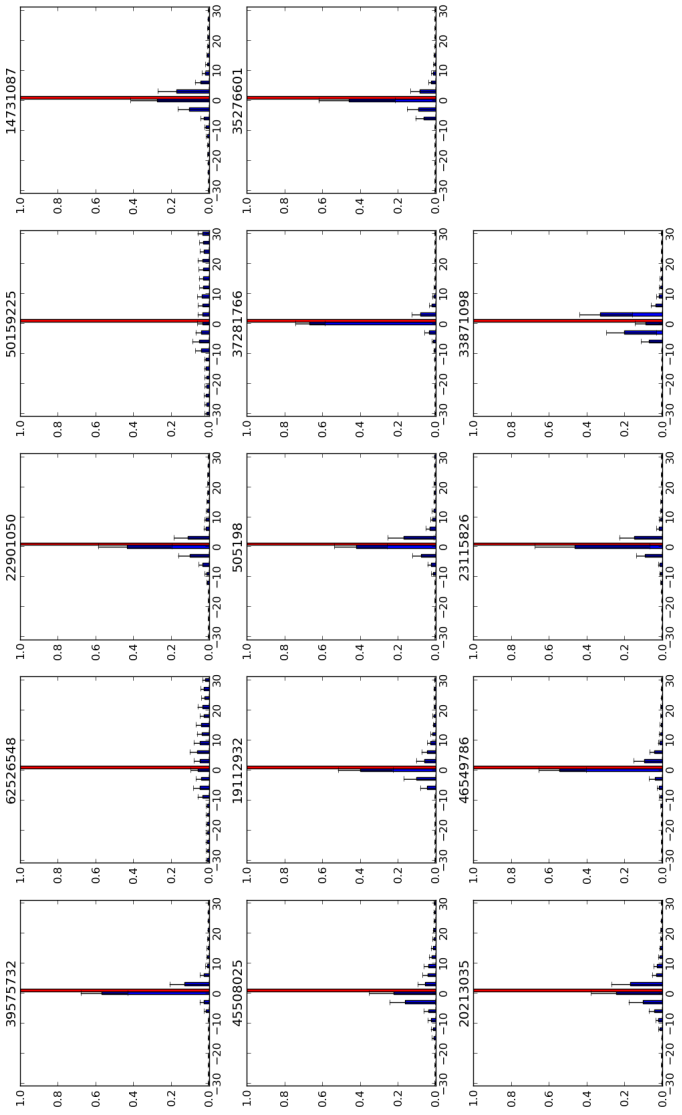


Figure 4.5: Allele frequency distribution prediction of alleles for the Gibbs sampler algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely of reference alleles based on a CHS population (red bars). Most of the predictions' allele frequency distributions center around the truth (reference). However at start positions 50159225 and 62526548, the posterior allele frequency distributions are close to the uniform prior distribution because there is little information from the data. They therefore would not create false positives as we had with the EM algorithm.

4.3.2.2 Two and three allele population frequency alleles

The next step in determining the efficacy of the two algorithms was to see how each performed when the allele frequencies were no longer all on one allele length, as well as not all allele lengths corresponding to the reference length. In determining this, we simulated two scenarios: first a two allele frequency distribution of ± 9 bp in the CLM population, and second a three allele frequency distribution in the CHS population with allele lengths corresponding to the reference allele, a -12 bp deletion and 6 bp insertion. Thirty loci at random were chosen in chromosome 20 for each of the two scenarios. Each algorithm then made calls at each locus whose resulting allele frequency distributions were scrutinized against the truth. Figures 4.6, 4.7, 4.8 and 4.9 illustrate the results of the two simulation scenarios for each algorithm.

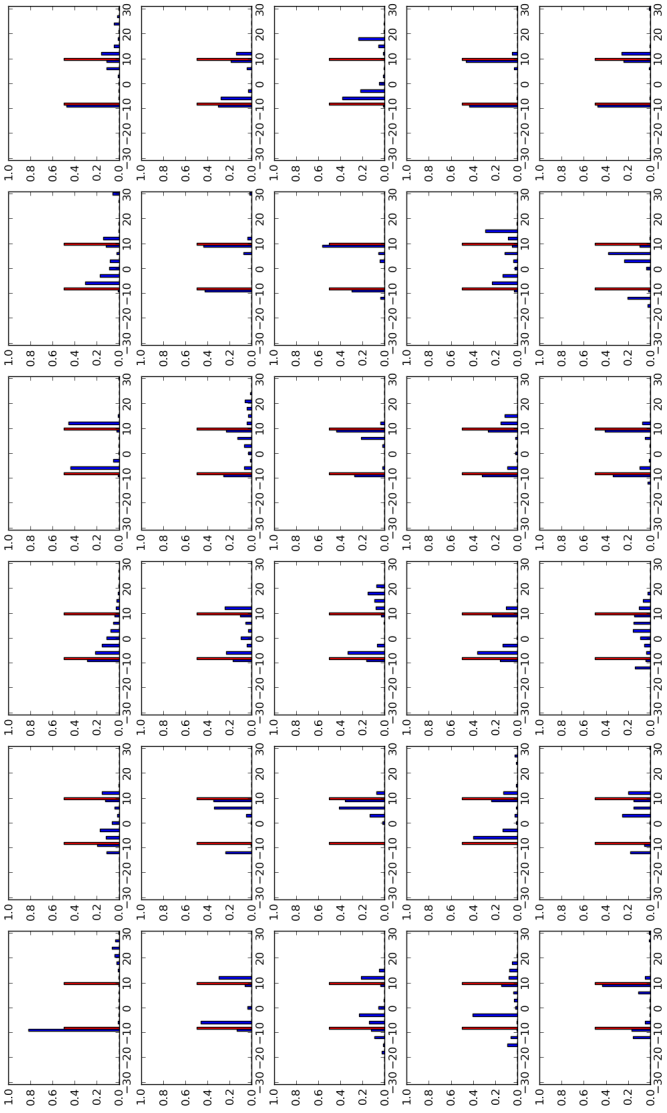


Figure 4.6: Allele frequency distribution prediction of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency distribution of ±9 bp each at a 0.5 frequency (red bars) based on a CLM population. As with the reference simulation, the EM is much more aggressive, yielding both stronger signals on the truth, as well as, overfitting at some loci, e.g. at the fourth locus in the bottom row.

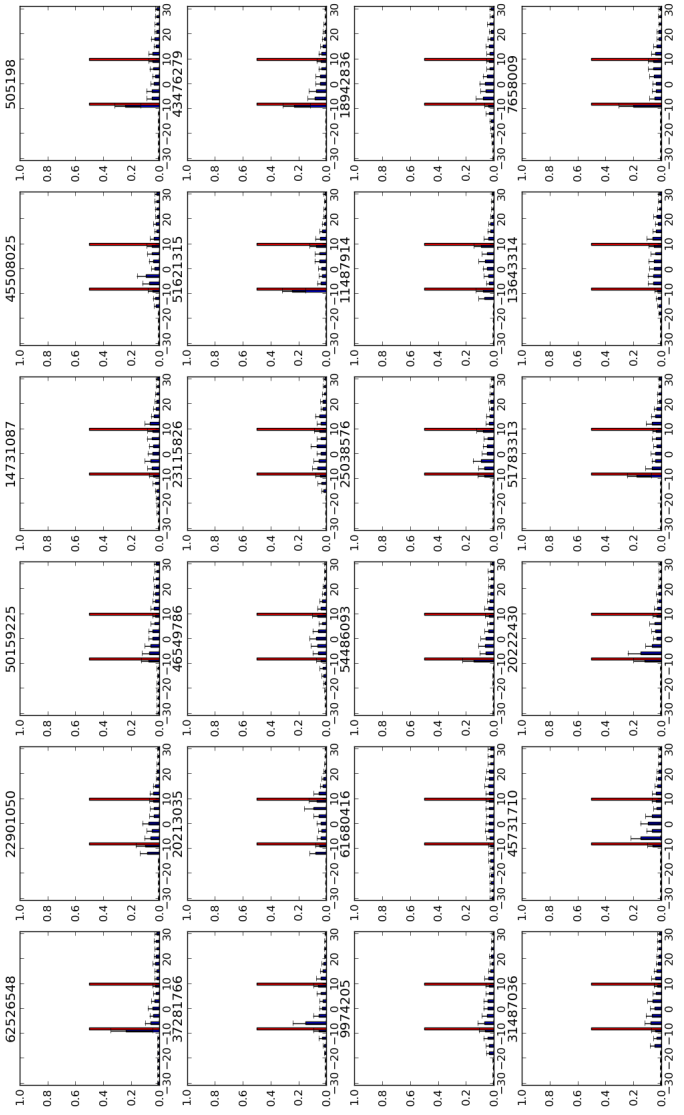


Figure 4.7: Allele frequency distribution prediction of alleles for the Gibbs sample algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of ±9 bp each at a 0.5 frequency (red bars) in a CLM population. Not as aggressive as the EM, sites show lower frequency peaks around the truth, but the Gibbs sampler, as before, does not overfit the data.

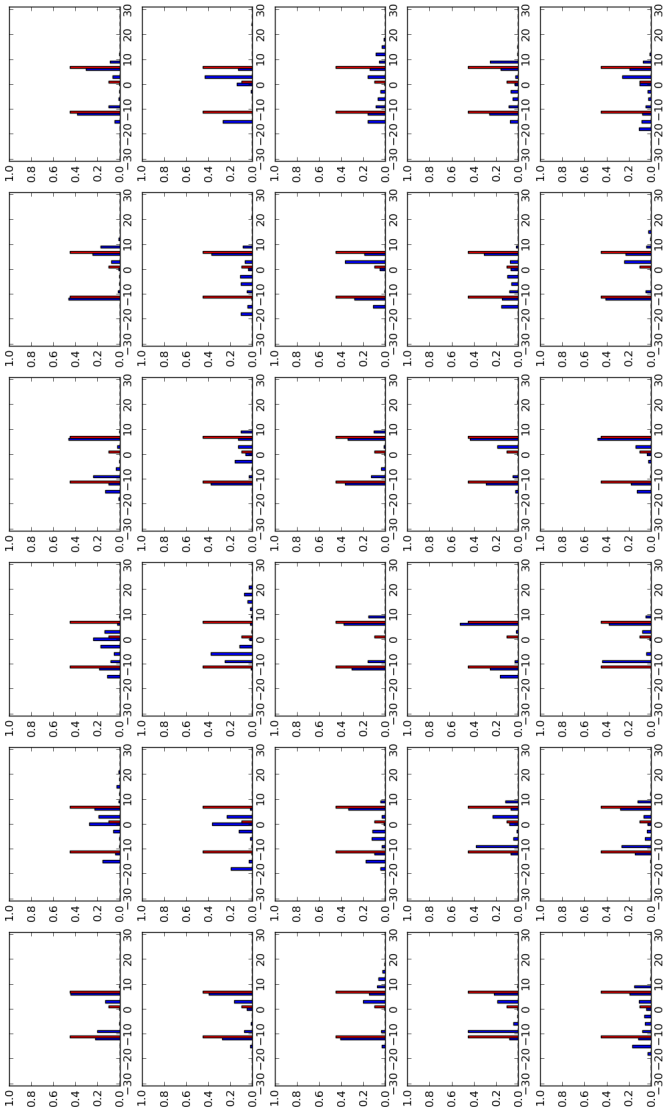


Figure 4.8: Allele frequency distribution predictions of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based on a CLM population.

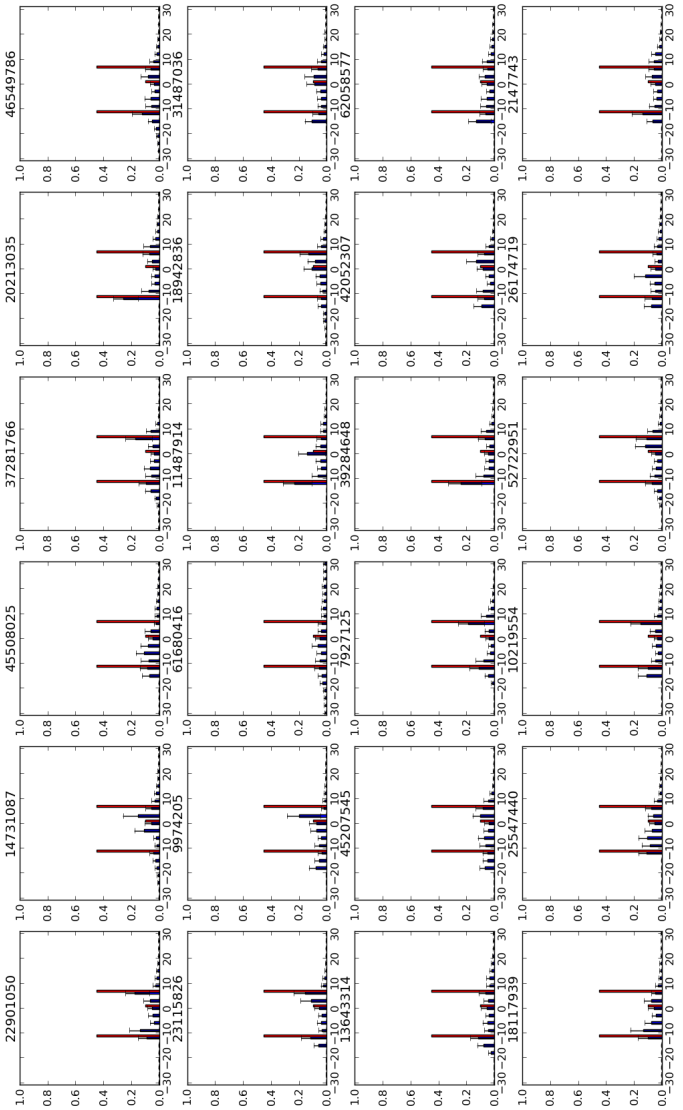


Figure 4.9: Allele frequency distribution predictions of alleles for the Gibbs sampler algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based on a CLM population.

4.3.3 Simulation results comparisons

After completing our three simulation runs, we sought to determine which algorithm worked the best, while yielding the fewest false positives. To start, we

looked at the average values each algorithm produced across all the loci for each of the simulation scenarios. This gave us an idea of how well in general the algorithms worked in ascertaining the underlying allele frequency distributions. Averages were found by amalgamating all the allele frequency vectors for each locus and then normalizing the values. The graph of these averages for each of the algorithms is shown in figures 4.10 and 4.11.

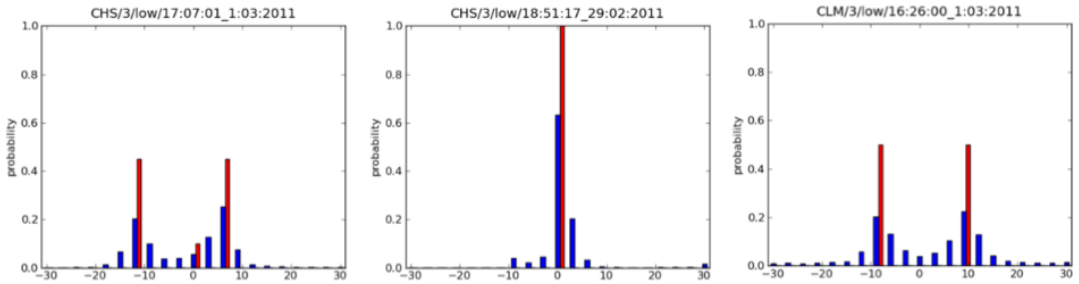


Figure 4.10: Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for the EM algorithm.

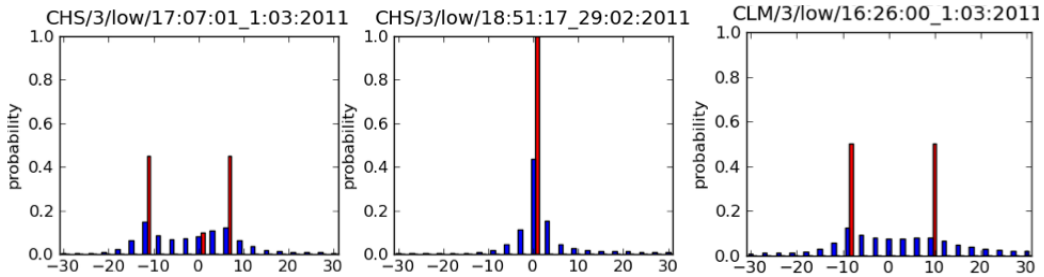


Figure 4.11: Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for Gibbs sampling algorithm.

Looking at the average frequency calls for both algorithms, it appears that both perform well under the reference scenario for non-reference calls, with neither method showing any systematic bias. It does stand to mention, however, that the EM algorithm is better at distinguishing between multiple alleles. In the two non-reference scenarios, the separation of allele frequencies is more clear cut for the EM than the Gibbs sampler. From this, it could be argued that the EM is a better choice.

However, aside from the overall averages of the allele frequency distributions for each algorithm, its important to look at a per locus accuracy rate as we are most interested in minimizing the number of false positive calls we make. As we have already noticed (see figure 4.4), the EM algorithm has a tendency of over fitting the data. When the amount of data is low – such that a putative repeat length is not observed – the EM forces all the weight onto a few allele sizes. When we plotted the values of the two algorithms on top of each other, it was clear that the Gibbs sampler, though not as conservative, didn't force the density onto a few calls. The Gibbs sampler also left some of the uncertainty intact while the EM did not. Figures 4.12, 4.13 and 4.14 show the comparison of the two algorithms against one another from a selection of the previously graphed loci above. The top graphs show where the EM predicts the underlying alleles accurately, and the bottom two graphs where the EM's predictions are overly aggressive.

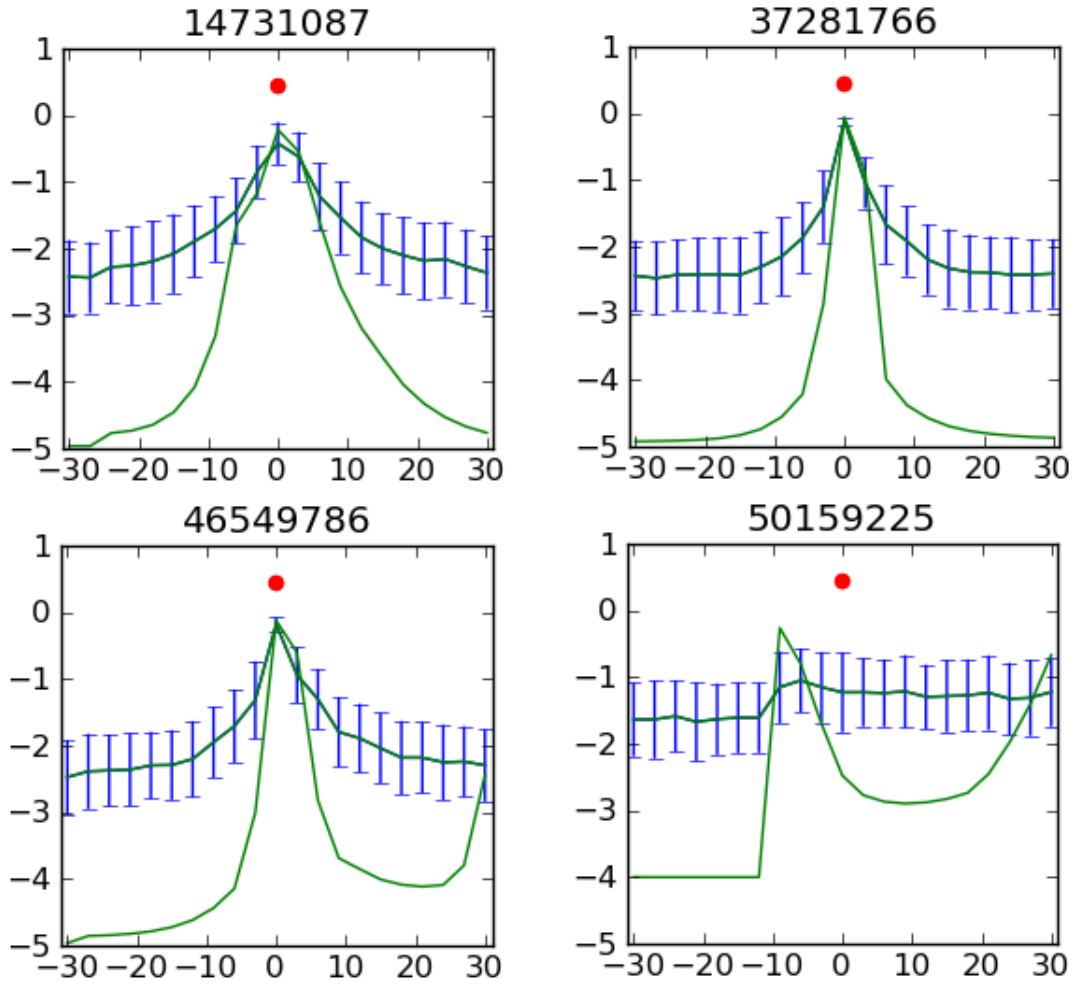


Figure 4.12: Comparison of the EM and Gibbs sampler algorithms for a reference allele frequency distribution. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying allele. The sole green line represents the values for the EM, while the green line with error bars represents the Gibbs sampler's predictions.

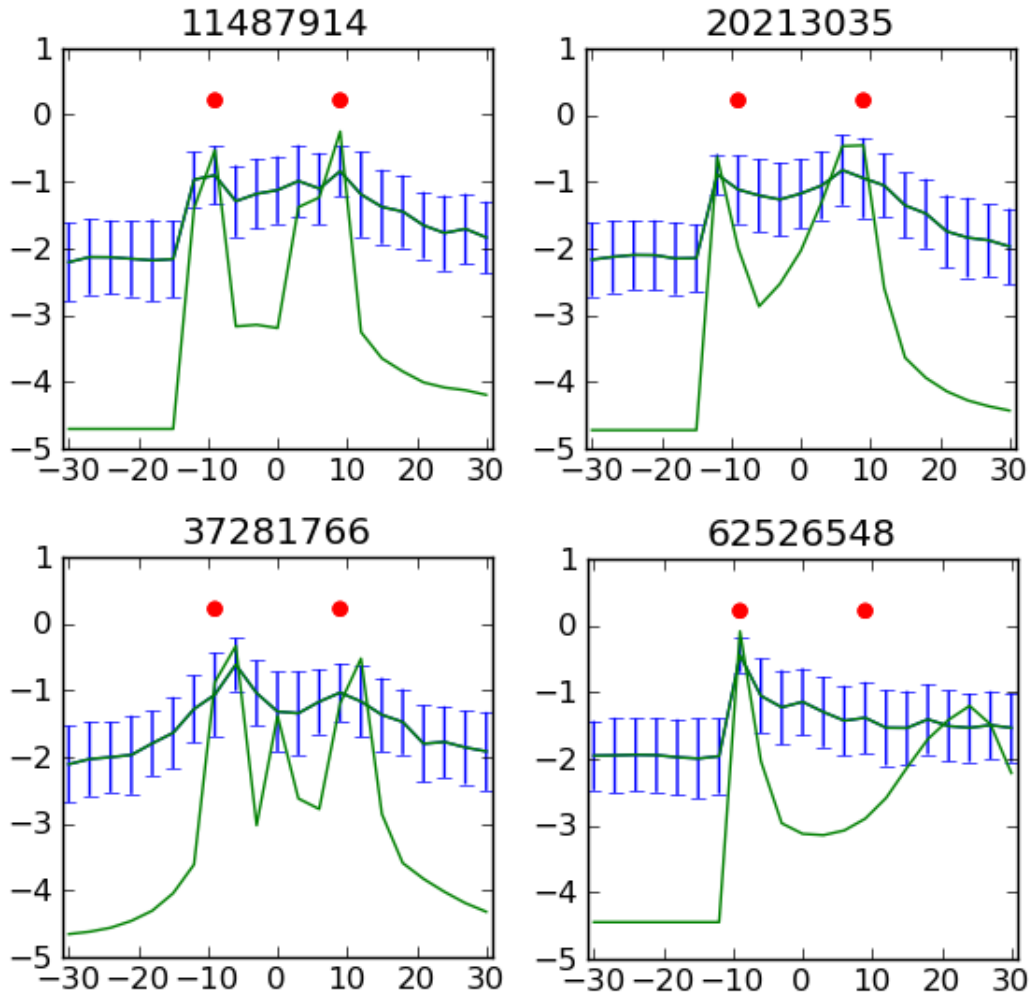


Figure 4.13: Comparison of the EM and Gibbs sampler algorithms for a two allele frequency simulation. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying alleles. The sole green line represents the values for the EM while the green line with error bars represents the Gibbs sampler's predictions. The top graphs show where the EM predicts the underlying alleles accurately and the bottom two graphs where the EM's predictions are overly aggressive.

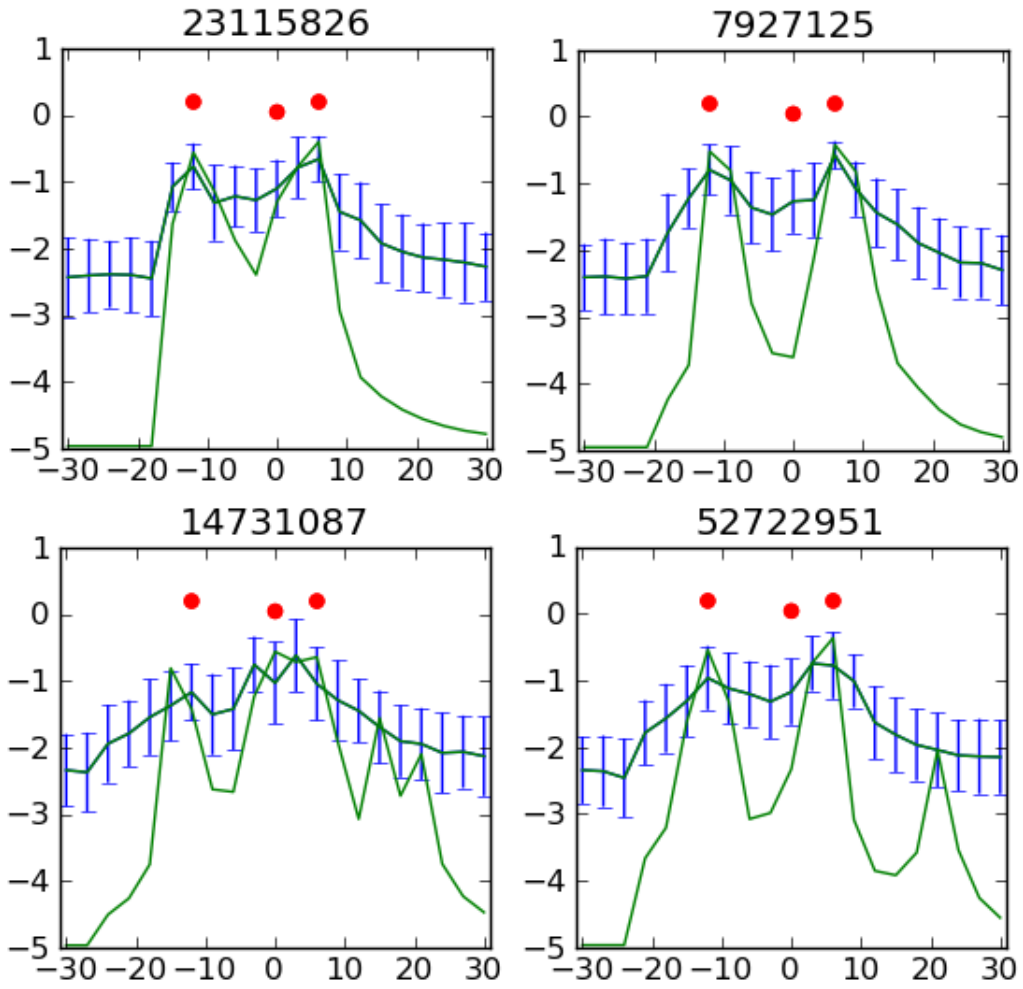


Figure 4.14: Comparison of the EM and Gibbs sampler algorithms for a three allele frequency simulation. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying alleles. The sole green line represents the values for the EM while the green line with error bars represents the Gibbs sampler's predictions. The top graphs show where the EM predicts the underlying alleles accurately and the bottom two graphs where the EM's predictions are overly aggressive.

Looking directly at the values of allele frequency distributions between the EM and the Gibbs sampler algorithms, it shows explicitly that the EM algorithm is much more aggressive compared to the Gibbs sampler and pushes almost all the weight into some number of alleles that it has evidence for. The EM algorithm does not follow the prior distribution (uniform in this case) when there is not enough data for an allele call, and therefore would cause many more false positives. Because of this, we used the more conservative Gibbs sampler for our analysis on real data.

4.3.4 Test statistics

When we try to gain inference from the allele vectors produced by the Gibbs sampler, it is important that we clearly define the statistics we wish to test so as not to obfuscate what the data is telling us. From the simulation results, which come from an idealized system, it does not seem plausible that we will make specific, single allele calls with the data at hand. The natural way to call specific alleles would be to set some threshold on the density and if an alleles density is above the threshold, we would claim that that allele is present in the population. Defining this value, however, would be difficult and would lead to either a large number of false positives or false negatives. An alternative approach is to look at the general composition of the allele frequency distributions. This line of thinking led us to calculate the entropy of the allele frequency distribution at a locus, as well as, how much of the density sits off the reference and ± 3 bp alleles.

4.3.4.1 Entropy

To begin, we shall first give the formal definition of entropy: the measure of disorder or unpredictability in a system. Mathematically, the entropy (H) of a discrete random variable X with possible values $\{x_1, ..., x_n\}$ (which for our system are allele lengths relative to the reference) is calculated as

$$H(X) = -\sum_{i=1}^{|X|} p(x_i) \log p(x_i)$$

where p is the probability mass function (amount of density on an allele) of random variable X. The base of the log can be of any value with the most common being e, 10 and 2 yielding the entropy in units of nats, dits and bits, respectively. It should also be noted that for values of $p_i = 0$ for any element i, the assigned value for the summand $0 \cdot \log 0$ will be taken as zero. In the context of our system, entropy is a measure of the amount of allele variability in our learned allele frequency distribution. Systems whose entropy are low means that the dispersion of data is also low (the true number of alleles is low). For instance, say at a particular locus, all the density was in a set allele on the reference: $p(reference) = 1$ and $p(allele) = 0$ for every other allele value. The entropy for this locus would therefore be zero. Now, assume that all the alleles are of equal frequency at that locus ($p(allele) = \frac{1}{21}$), the entropy would then be 1.322 (in base 10). This scenario would represent the maximum entropy for an allele frequency distribution. An allele frequency distribution which predicts a multiallelic locus would have a high entropy, while a locus that has most of its density on a specific allele would have a low entropy. Explicitly, this statistic would declare which loci are actively evolving or have a large number of alleles at a locus. While a locus with a high entropy doesn't tell us much about the actual allele frequencies other than that they vary more than a low entropy locus, hypothetically a low entropy locus would give us information we can use to determine whether the set allele(s) is on the reference or not. To do this, we need to look at how much of the allele density is off the reference/± 3 bp.

4.3.4.2 Off reference/± 3 bp

We consider two different statistics to measure whether the density away from the reference is sufficient to say that there are non-reference alleles within the population at that particular locus. Both these statistics are calculated simply by subtracting either the learned frequency of the reference allele from one, or the sum of allele frequencies of allele lengths $+3, 0, -3$ bp from one. Ideally, we would be able to use one of these statistics in concert with the entropy statistic, and from this, be able to tell a lot more about the locus than by each statistic separately. For a locus which has a low entropy value but a high density off

reference/±3 bp, we would believe that there is most likely a set allele at that locus that does not coincide with the reference. However, as we will see below, having low entropy and a high on reference density act as the null values for our testing whether or not a statistic's value at a locus is significant enough to assign a call to it. This makes inference in the opposite direction more difficult.

4.3.5 False discovery rate

To accurately attribute some categorical value (actively evolving, off reference) to each locus within a population (as described in 4.3.4.1 and 4.3.4.2), it was important to first determine what values were in fact significant and which ones weren't. This was accomplished by extending our reference simulation to all triplet tandem repeat loci (1,881) on chromosome 20 for each population. This yielded 26,334 (1,881 loci · 14 populations) allele frequency distributions. Using the methods described in 4.3.4.1 and 4.3.4.2, we calculated the values for entropy, off reference and off ±3 bp for each locus in each population. As we know that each of these sites were simulated under the condition that every allele for every individual for every locus matched the reference length, we were able to calculate the false discovery rate (FDR) at a given cutoff (c) for each population as follows

$$FDR = \frac{\sum_L \mathbb{I}[s_l > c]}{|L|}$$

where L is a set of loci and s_l is the statistic value being tested (entropy, off reference/±3 bp). For entropy, we iterated through cutoffs ranging from [0,2.5] by increments of 0.001, and for the off reference/±3 bp, iterated through cutoffs ranging from [0,1] by increments of 0.001. This ultimately yielded a full range of FDR values from 0 to 1 and the associated cutoff value for each FDR value.

We applied the methods described above to all 1,881 triplet repeat STRs on chromosome 20 for all 14 populations, using each of the test statistics and both the conservative and decay priors. This makes $1,881 \cdot 14 \cdot 3 \cdot 2 = 158,004$ tests in total.

Next, for each cutoff threshold we subtracted the number of false positive calls we

would expect to observe based on our FDR simulations, and plotted the net estimated number of true calls against the FDR. We refer to true calls as the number of loci called whose value is above the cutoff minus the number of expected false positives. For example, if in the real data we observed 400 sites which are above the cutoff for a FDR of 0.05 (chosen to minimize the number of false positive calls), this means that out of all these 400 calls, roughly 94 are false positives $(1,881 \cdot 0.05)$. Taking these false positives into consideration, we are left with 306 true calls (400 - 94). Shown below are the plots for each statistic/prior pair for three different populations.

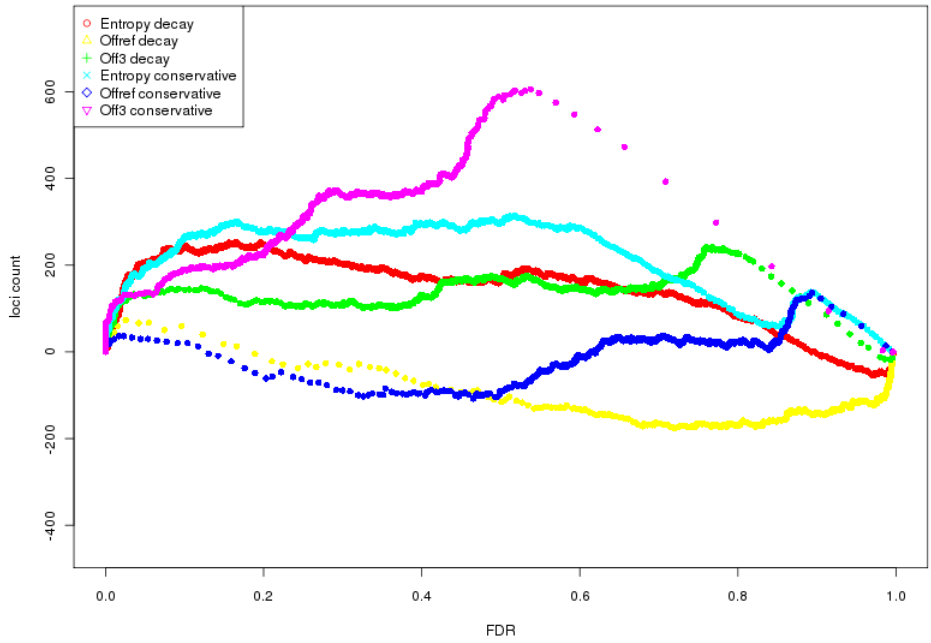


Figure 4.15: Plot of FDR versus true calls for the ASW population for triplet repeat loci on chromosome 20.

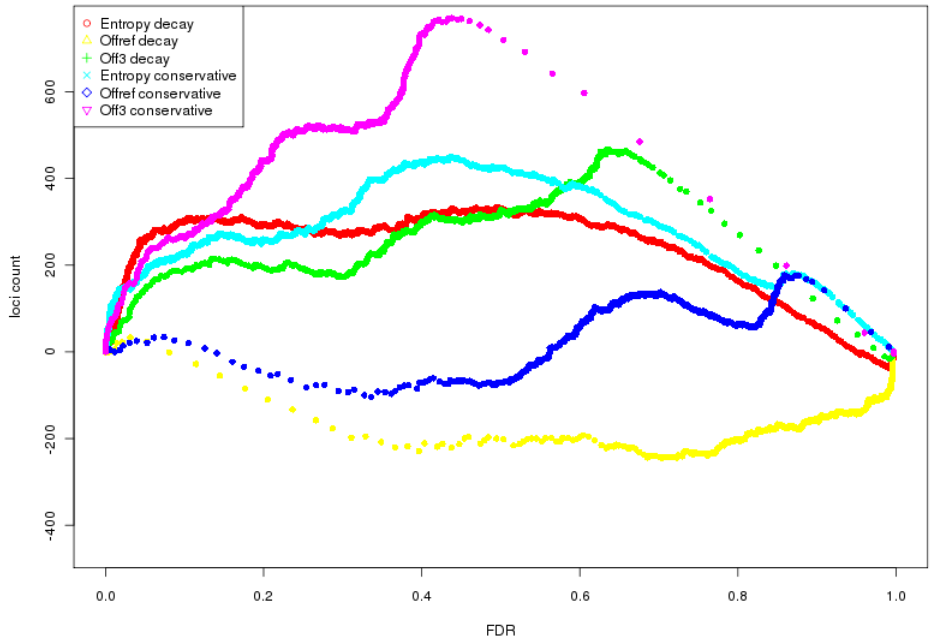


Figure 4.16: Plot of FDR versus true calls for the MXL population for triplet repeat loci on chromosome 20.

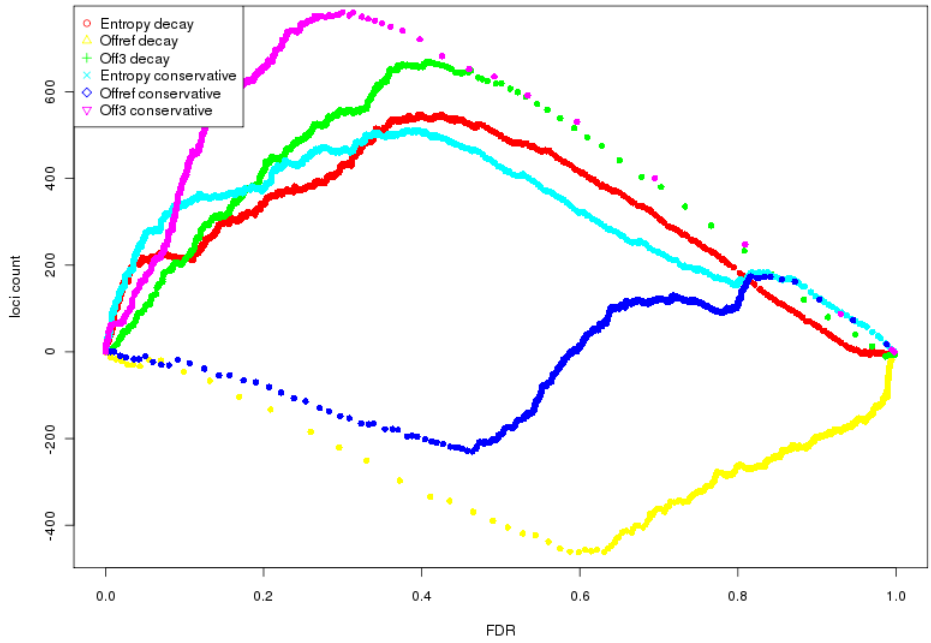


Figure 4.17: Plot of FDR versus true calls for the PUR population for triplet repeat loci on chromosome 20.

These plots (4.15, 4.16 and 4.17) show a clear advantage in the number of true calls for the statistics entropy and off ±3 bp. At a FDR of 0.05, the average weight off ±3 bp for all populations using the decay prior is 0.966 (range of [0.952,0.977]) and 0.951 (range of [0.915,0.969]) for the conservative prior. We chose to exclude population IBS as it was only sequenced from six individuals and its calculated off ±3 bp weights were 0.765 and 0.463 for the decay and conservative prior, respectively. The number of loci above the cutoff at a FDR of 0.05 for both entropy and off ±3 statistics using both priors is roughly 90 calls for each population. Therefore, given our analysis is only on chromosome 20 and assuming it is representative of the rest of the genome's ratio of significant loci to non-significant loci, we would expect to observe over 4,100 independent loci with significant values for each of the statistic/prior pairs.

We also observed at a number of FDR values (particularly in the off reference statistic) whose number of expected true calls were negative. This could be be-

cause the real data is subject to reads not mapping uniformly around real sites (as they did in our simulation), so the MPERS observed don't actually come from the genome wide MPERS distribution. It may also come from multiple low frequency alleles in the population whose frequencies' are not large enough to be picked up by the Gibbs sampler, and are therefore washed away by the prior, making reference calls more likely.

4.4 Results

We marked out loci across all populations that passed a cutoff corresponding to a FDR of 0.05 by combining the calls made with either prior. The highest number of significant loci coming from the combined prior calls was made by the entropy statistic (1,361 unique loci) followed by the off ± 3 bp statistic (1,019 unique loci) and lastly the off reference statistic (733 unique loci). The number of calls per prior were almost equal: 1,609 unique loci coming from the decay prior and 1,617 unique loci coming form the conservative prior. From here on, we shall focus our analysis on the entropy and off ± 3 bp statistics.

We next looked at how many loci are called in multiple populations (≥ 5) for the same statistic (entropy and off ± 3 bp) and diagrammed the intersection of the two statistics' calls (see figure 4.18).

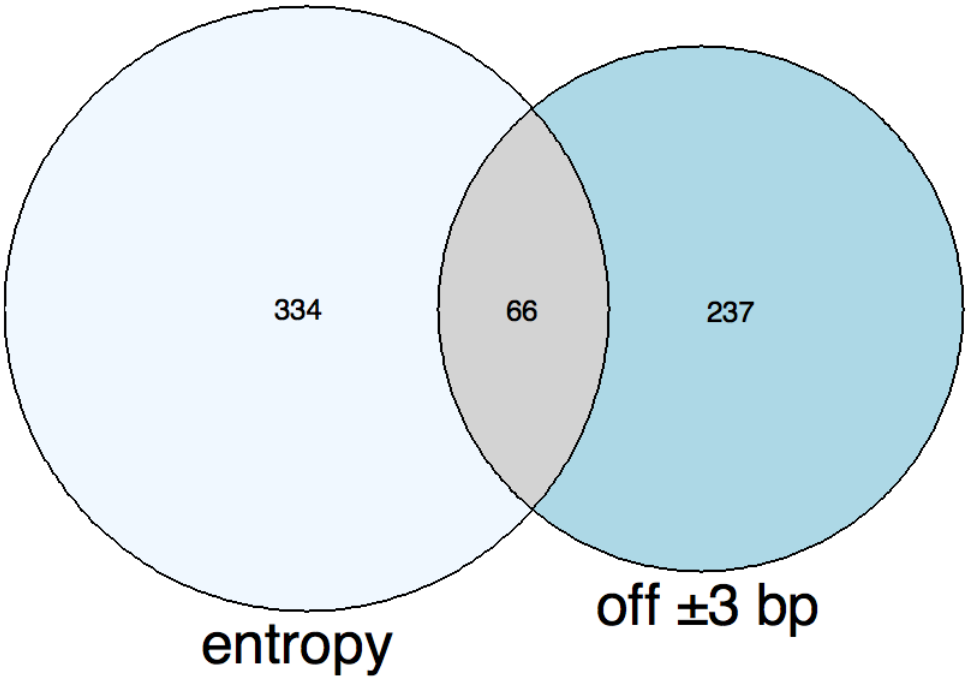


Figure 4.18: Venn diagram of intersection of significant loci called by entropy and off ±3 bp.

4.5 Discussion

For sites where there is a trend for the off ±3 bp statistic in multiple populations, it most likely means that the reference is the minority global allele (303 loci having a call for off ±3 bp statistic in five or more populations). Loci which have calls for the entropy statistic in multiple populations mean that these loci are more likely to be actively evolving and less likely to be under selection (400 loci having a call for entropy in five or more populations). On the other hand, its harder to say which sites are truly reference or under selection as these values represent the null in our modeling.

When we looked for loci which were called both for entropy and off ±3 bp, we found that only 66 sites matched this criteria. This is not altogether that surprising. These results are consistent with it being unlikely for there to be a dispersed distribution of allele sizes but almost no reference allele. One would

expect an actively evolving site to contain at least some density on the reference allele length in the population.

4.5.1 Factors

As an extension to our analysis in chapter 3 of how the factors of a repeat locus affect the probability of observing an indel, we decided to explore the same factors as described in chapter 3 for our two population statistics. To begin, we first fit a logistical model on whether or not a locus was called using criteria for entropy and off ± 3 bp statistics (at an FDR of 0.05). We next fit a linear model for sites which were called significant and explored how the factors affected the value of the two statistics. The values were modeled independent of which prior they came from; meaning all calls for both priors were lumped together. The plots for coefficient values are shown below in figures 4.21, 4.22, 4.21 and 4.22.

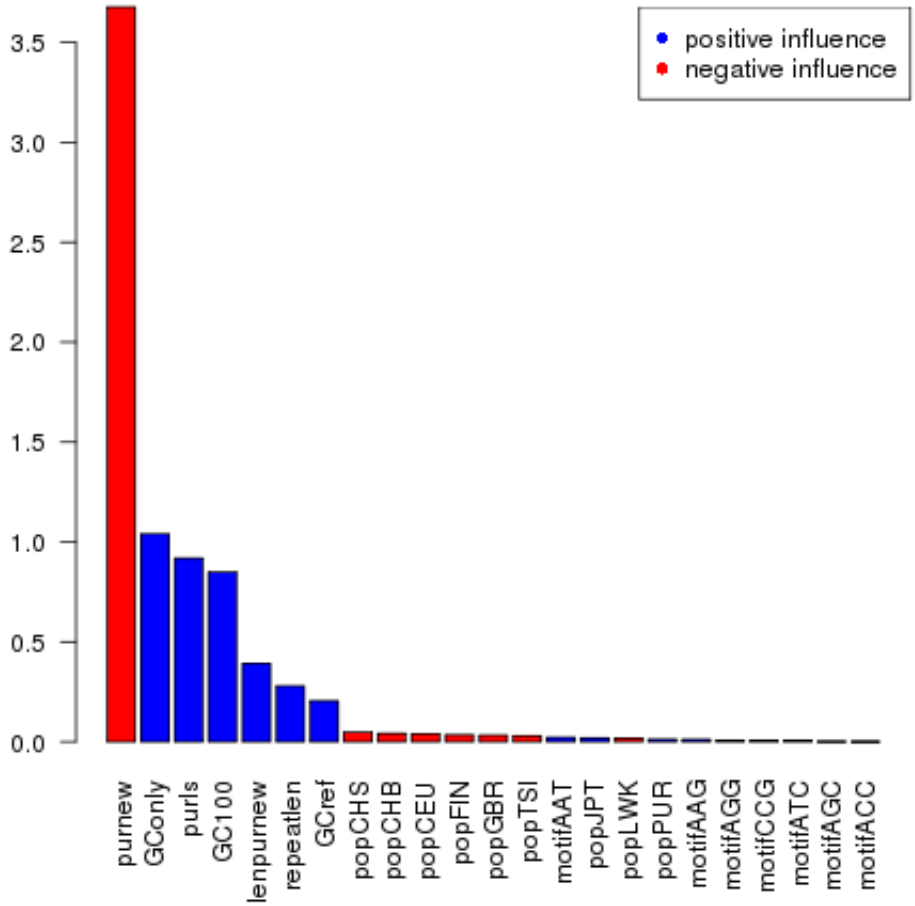


Figure 4.19: Bar graph of absolute values of coefficients from logistic linear model on whether a locus's entropy value is significant against various factors.

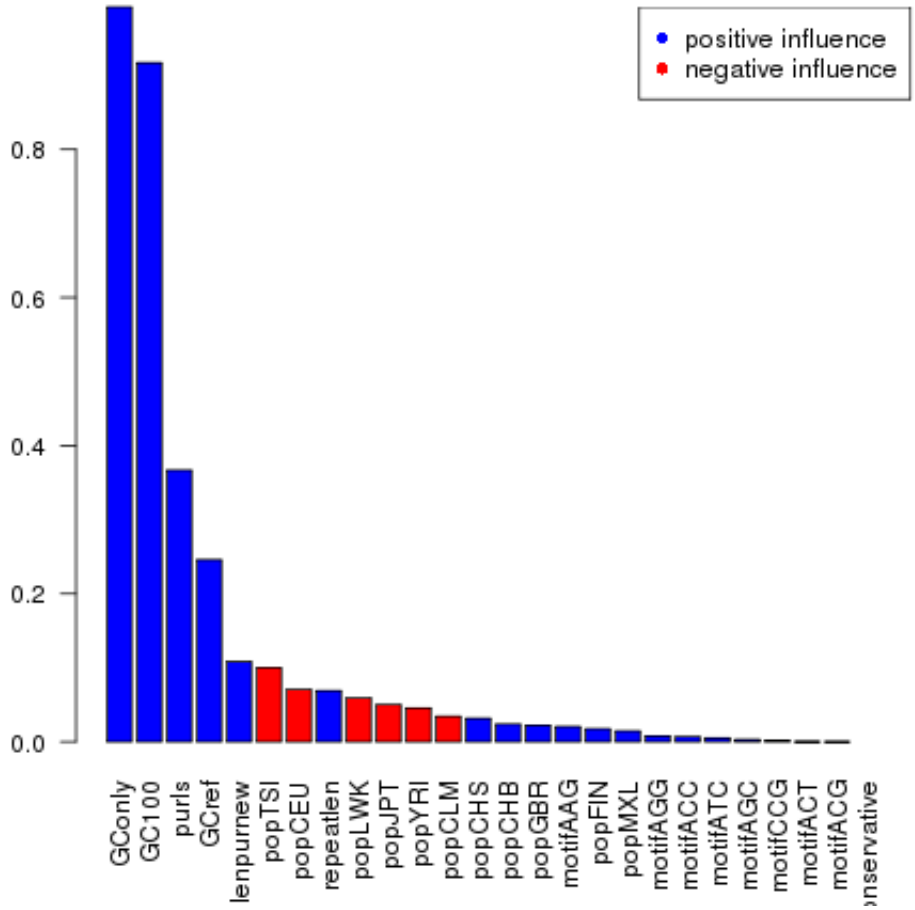


Figure 4.20: Bar graph of absolute values of coefficients from logistic linear model on whether a locus's off ±3 bp value is significant against various factors.

First looking at the logistic modeling of whether a locus has a significant value for both the entropy and off ±3 bp statistic, we observe that many of the factors values seem to be relatively in the same order of significance, direction and magnitude. The statistic GConly has the strongest influence on a locus having a significant value for both statistics followed closely by both purity and GC content statistics. The population factors are relatively insignificant for the entropy statistic and have some influence in the off ±3 bp statistic. In the off ±3 bp statistic, the strongest correlations are negative (compared to the ASW population) in

populations TSI, CEU, LWK, JPT, YRI and CLM. Inspection of these populations' sequencing statistics gives no reason as to why some populations might be more readily called than others. Furthermore, CHS and CHB (two closely related populations) have relatively equal correlations in the same direction. This would lead us to believe that there might truly be correlations in populations which warrant further inspection. The motifs have relatively little influence, with AAG having the strongest correlation (positive) which is exactly the same as observed in our chapter 3 results. The only motif with a stronger signal in the previous chapter's modeling was that of AAT (which had a low p-value in our modeling and was therefore not graphed). The prior had no influence on the system.

If we now go back and scrutinize the larger coefficient values with those in the logistic linear models in chapter 3, the coefficients are at relatively the same value and rank, however, GConly and GC100 are both negatively correlated with observing a variant when they are positively correlated with having significant values for entropy and off ±3 bp statistic. While both populations YRI and CEU (from which the individuals in chapter 3 belong to) are negatively correlated with the entropy and off ±3 statistics, this most likely doesn't account for this reversal in influence. Another explanation could be that while the 1000 Genomes Project's individuals are sequenced to a lower depth, their combined reads are enough to overcome the bias in less reads mapping to loci whose proximal sequence is GC rich (see chapter 3). However, the strongest explanation requires us to think back to the values of the of the linear regression for magnitudes of indels in chapter 3. The values for this model showed that the GC content was positively correlated to there being larger indels when they were observed. Allele frequency distributions which have smaller alleles would most likely not have enough power to be called from our entropy and off ±3 bp tests. This knowledge indicates why the larger indels (which would give rise to higher entropies and off ±3 bp values) would be positively correlated to the amount of GC content in a region, as observed previously.

We next fit a linear model to the values of both statistics conditioned on the statistic's value being significant.

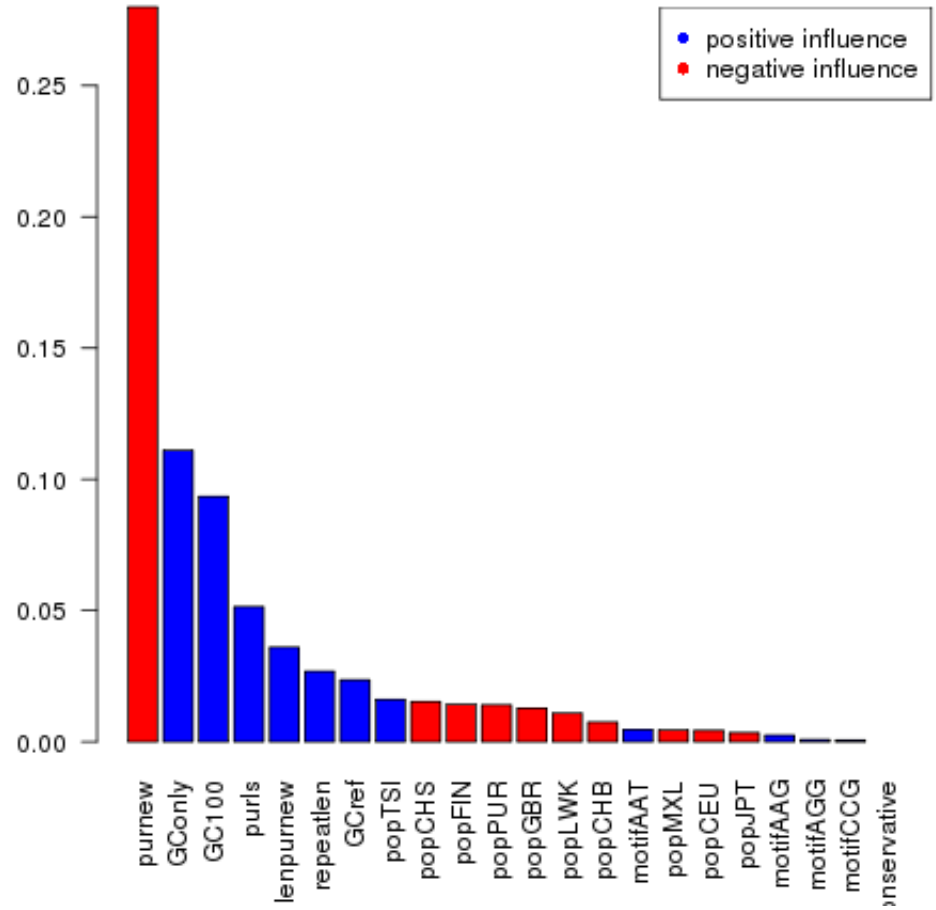


Figure 4.21: Bar graph of absolute values of coefficients from linear model of significant entropy loci values and the various explanatory factors.

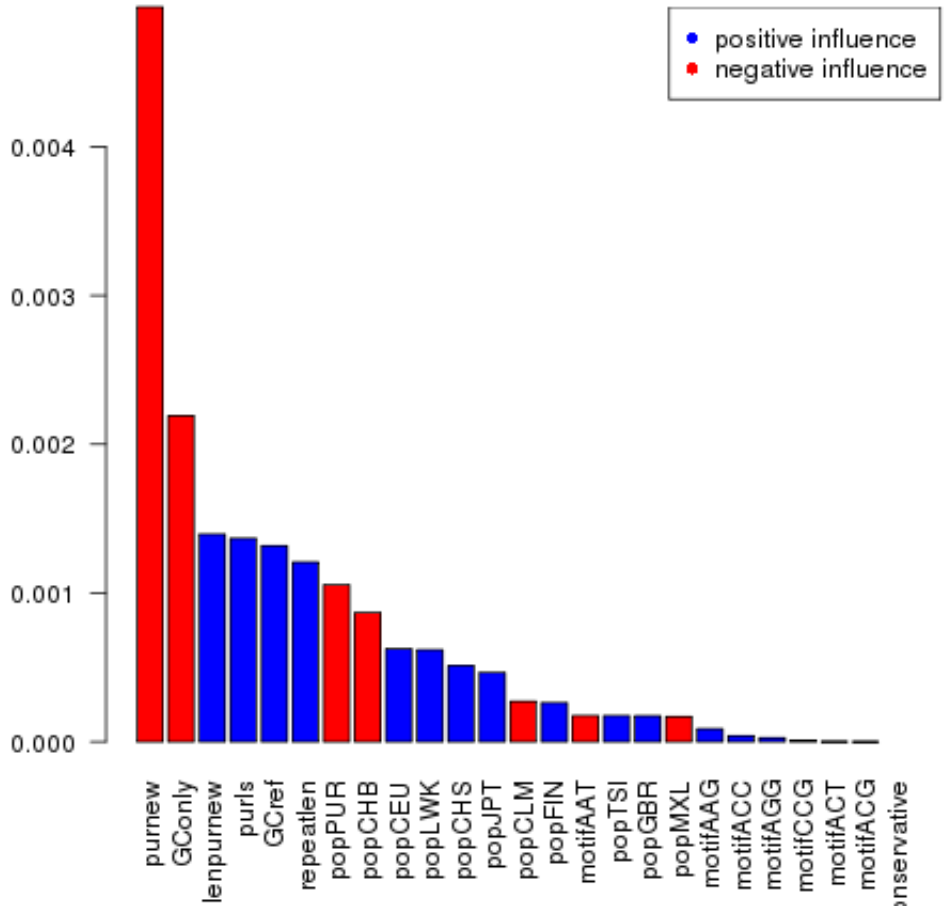


Figure 4.22: Bar graph of absolute values of coefficients from linear model of significant off ±3 bp loci values and the various explanatory factors.

The same trend in relative size and order is observed for both statistics as was seen in the previous logistic regression. The only difference being, for the off ±3 bp statistic, GConly negatively influences higher off ±3 bp values. This reversal is most likely an artifact of the low number of sites used to fit this model – as seen by the extremely small coefficient values. The values are further corroborated by comparison to the linear regression for the magnitude of indels in chapter 3 which shows an almost identical order and relative influence of one factor compared to another.

The main hinderance in the modeling of this call set is that the number of loci assayed is low. However, it is encouraging that the modeling of factors in this chapter and the previous chapter corroborated, which leads us to believe that the results are correct and that the learned coefficients do in fact correctly model the influence each factor has on the allele frequency distribution of tandem repeat loci.

4.6 Conclusion

Inevitably, our power to model and make inference in this system comes down to the number of individuals sequenced in a population and their combined sequencing depth. For the 1000 Genomes Project data set, split into populations, it would appear that there is enough data to give some relevant information about the tendency for a site to be variable, but nowhere close to enough read information to determine the exact frequency of each allele in a population. A further study could look back at the reported allele frequency distributions and make predictions on a range of alleles by setting some threshold on the amount of density needed to attribute a specific variant in the population. A good starting place would be places where there is significant weight in the off ± 3 bp statistic. One approach to get more information will be to combine the populations into a global population and see how this affects the values of the statistics at each locus. We presume that loci that were found to have calls shared across all populations will continue to be found in this joint analysis, and we also believe that amalgamating the data might also give enough information to call loci which previously went uncalled in the individual populations. We have not been able to carry out this combined analysis yet because of compute resource limitations in our implementation.

We modeled the effect each factor (as described in chapter 3) has on the values of our two statistics in both a logistic linear and linear model. The values of coefficients we found from the modeling were in line with the values and direction

of coefficients we had observed in the previous chapter – and when not, a explanation was presented as to the cause of the discrepancy and therefore explained away in context. The continuity of coefficients between the two chapters illustrates the viability of this type of exploration in tandem repeat loci. Further, as the 1000 Genomes Project data set grows, we believe our exploration using this data will broaden our understanding of what role each factor plays – and to what extent – in the variation of tandem repeats.

Chapter 5

Conclusions

5.1 Conclusions, discussion and future work

For the past four years, I have endeavored to understand a specific area of genomic diversity that warrants attention. STR loci remain difficult to type using new sequencing technology, and because of this, are not fully characterized. My research has sought to produce a reliable model to type the STR loci of high sequencing depth individuals using paired end read next generation sequencing data. From these calls, I sought to characterize the factors which increase or decrease the probability of observing a variant at a locus. My single sample variant calling model was then reformulated to look at the overall genomic diversity of STR loci in population data of low sequencing depth individuals.

5.1.1 Modeling variation in STRs

The development of STRYPE (chapter 2) has added a new tool to the genomic variation community that has been specifically designed to type STRs. Because of their variability within a population, being able to type STRs will assist in both evolutionary and disease analysis. More so, as many triplet repeats are associated with – or even the causative factor of – many diseases, additional typing of STRs may lead to further discoveries.

However, as sequenced reads become longer (from 35 to upwards of 100 bp as

discussed in the introductory chapter) the number of sites in the genome which are unable to be typed by split alignment start to diminish quickly. This does not, in fact, remove the need for alternative methods for typing short tandem repeats. Split alignment algorithms will ultimately be the standard indel callers as technology develops, but they are still constrained by the length of the read and computational limitations. Longer tandem repeats will still remain unassayed, as well as, larger indels which are prohibitively expensive to explore due to the large computational requirements needed to accurately determine the size of an indel. The latter – and I believe larger problem – will remain the limiting factor until computational power increases to a point where large scale searches of indels of varying sizes are not longer prohibitively expensive. That being said, there exists huge amounts of data that has been sequenced with shorter reads, and to low enough depths, that normal split alignment tools may be unable to correctly type – from the three trio families and the 1000 Genomes Project described in this report to the UK10K Project, as well as, many non-human genomes and projects (1001 Genomes (*Arabidopsis thaliana*) project). Without the methods described here, untold magnitudes of variation would be overlooked. As these sequencing projects are already underway or complete, to maximize the benefit of their sequencing, it is important that the largest amount of information be gleaned from this sequencing and we believe our method does just that.

5.1.1.1 Future work

In modeling variation in a deep sequenced individual (using a Bayesian approach), we needed to describe a prior which mitigated the problem of over fitting a sample's sequencing data to our calls (described in chapter 2). For our original implementation of STRYPE, the prior was based solely on the calls made from low sequencing depth capillary reads which only gave us the probability of observing a single allele of a given repeat length compared to the length of the STR in the reference. An additional heuristic prior was later added to correct overcalling of less likely genotypes. Since we are now able to type more and more STRs across multiple individuals, we can exploit the resulting information by feeding it back into our model. From this, the validation and simulation data can be

applied in tandem to learn the true genotype and indel magnitude prior, yielding a more descriptive and biologically accurate prior we were without in our initial modeling.

5.1.2 Characterizing STR variation

The analysis of influences different factors have on a STR exhibiting variation has been limited by the ability of researchers to type many STR loci across many individuals from a single sequencing platform. Though having its own limitations, whole genome wide shotgun sequencing using next generation sequencing machines has given geneticists access to magnitudes more sequence data.

As STRs can be characterized by a relatively small number of factors, it is possible to learn the influence each factor has if a sufficient number of loci are typed across multiple individuals. Doing just that, we were able to determine the influence a variety of factors have on triplet repeat STR variation by typing nine deeply sequenced individuals and regressing the factors against both the observation of a variant and the size of the variant.

5.1.2.1 Future work

As we focused solely on triplet repeats, the natural progression will be to broaden our assay to all STRs. To this end, we have identified, using TRF, all 1-10, 15 and 20 bp motifs. It will be interesting to see how the various factors influence variation across different motif sizes. We presume the length of the longest pure stretch will remain the strongest influence, but whether the other factors remain relatively the same will be an interesting study. However, we should point out that triplet repeats (the focus of this report) are a special set of tandem repeats within the genome and may not be representative of tandem repeat polymorphism in the human genome. As discussed in the introductory chapter, the absolute number of triplet repeats in the human genome is not in line with the number of loci you'd expect to observe given the trend of decreasing number of loci as motif length increases. This gives credence to the belief that these sites are different

and may behave differently when undergoing mutation than the other tandem repeats, a consideration to keep in mind when modeling the different factors that effect the probability of observing an indel at a given repeat locus. Also, since the triplet repeats' motif length is the same as a reading frame during translation, this would most likely cause them to act much differently in transcripts – especially in exons – than other tandem repeat motif lengths. It is also clear from other indel callers, such as DiNDEL, that different motif length tandem repeats exhibit different characteristics, as is the case for homopolymers. While homopolymers are more likely to exhibit sequencing errors, some tandem repeats may fold back on themselves which could introduce a bias during sequencing (such as the intrastrand hairpin structures formed by the CAG/CTG class of triplet repeats which have been associated with neurological diseases). All these considerations should be explored and modeled in future implementations. And lastly, once we have ascertained the relative influences of each factor, to compare them across all motif lengths would be of great interest. Comparing them side by side would illustrate what effect (if any) the length of the motif plays in STR variation.

5.1.3 Modeling STR loci in large population data sets

Understanding and defining population scale genomic variation is at the forefront of bioinformatics research. The low cost and rapid pace of sequencing of whole genomes has made it possible for geneticists to describe genomic variation down to allele frequencies of a percent or below in a population for SNPs and small indels. In hand with this, we sought to understand STR variation on a population level by calculating the entropy and off reference/$\pm$3 bp weight of variants at a given locus in a population. This study provided a set of loci on chromosome 20 that were shown to be either more variable than expected or whose distribution of variants at a locus is not best described by the reference.

5.1.3.1 Future work

Memory restraints limited our prototyping to chromosome 20 as it was relatively easy to assay due to its size (about 2% of the whole genome). However, we

were unable to run a global data set with all the individuals of each population combined together. As each population is comprised of a different number of libraries – that are all in turn sequenced to a different depth – its best to start by generalizing when discussing the constraints in running a full population assay. So to start, the number of libraries per population ranges from 6 to 143. For a single population, our program loads each likelihood file into memory (ranging in size from 3,852 to 7,340 Kbs, with a mean and standard deviation of 6,640 and 829 Kbs, respectively), and then models the most likely configuration of allele lengths at a given locus in a population (as outlined in chapter 4). Given the population with the largest number of libraries (CHB), the amount of memory needed just to read in the files (at an average size of 6.6 Mbs) – and excluding all the overhead – is roughly a GB. With Perl's overhead, this brings the memory requirement to just under 2 GBs – which is the maximum allotted memory for a job to run without explicitly requesting more memory. However, on a good note, the computational requirements were well within the limits; the longest run (as we ran each population a few times to assure that the sampling was working) took 70679.96 CPU seconds. When we tried to run the global population by combining all the libraries (1173), this brought the baseline memory requirements to 7.7 Gbs – not including the overhead. This pushed our memory requirements well above the standard memory allotment. Because of this, the next step will be to figure out a way to reformulate our model so that we can run both full genome and global data sets or request a much larger segment of memory to be allocated to our population run – an expensive and somewhat wasteful proposition. The more sensible, albeit difficult and time consuming task, would be to rework the model such that only one locus is read in at a time. This does have the consequence of taking much more time computationally but we must weigh out the cost and benefits of using up more memory or more computational cycles – a question best posed to the system's administrator of our supercomputer farm.

Ultimately, as chromosome 20 had relatively few triplet repeat loci (1,881) compared to the rest of the genome (80,868 in the autosomes), we are sure that many more sites will be found which warrant attention and whose factors can also be

scrutinized as those in the high depth sequenced individuals. And as before, we will be able to consider the motifs of all lengths as well.

References

Albers, C., Lunter, G., MacArthur, D., McVean, G., et al. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21(6):961, 2011. 4

Ananda, G., Chiaromonte, F., and Makova, K. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biology*, 12(3):R27, 2011. 87

Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, 9(13):3015, 1981. 37

Ball, E., Stenson, P., Abeysinghe, S., Krawczak, M., et al. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Human muttaion*, 26(3):205–213, 2005. 10

Bansal, V. and Libiger, O. A probabilistic method for the detection and genotyping of small indels from population-scale sequence data. *Bioinformatics*, 2011. 25

Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2):573, 1999. 6, 29

Bentley, D., Balasubramanian, S., Swerdlow, H., Smith, G., et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. 4, 16, 28, 31

Bhangale, T., Rieder, M., Livingston, R., and Nickerson, D. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. *Human molecular genetics*, 14(1):59, 2005. 25

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., et al. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics*, 62(6):1408–1415, 1998. 87

Brownlee, G., Sanger, F., and Barrell, B. Nucleotide sequence of 5S-ribosomal RNA from Escherichia coli. *Nature*, 1967. 3

Bugaut, A. and Balasubramanian, S. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, 47(2):689–697, 2008. 115

Calafell, F., Shuster, A., Speed, W., Kidd, J., et al. Short tandem repeat polymorphism evolution in humans. *European Journal of Human Genetics*, 6(1):38–49, 1998. 87

Carrilho, E. DNA sequencing by capillary array electrophoresis and microfabricated array systems. *Electrophoresis*, 21(1):55–65, 2000. 3

Chen, K., McLellan, M., Ding, L., Wendl, M., et al. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome research*, 17(5):659, 2007. 5

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., et al. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic acids research*, 37(suppl 1):D19, 2009. 28

Cohen, J. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biology*, 2(12):e439, 2004. 2

Consortium, T..G.P. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010. 31, 72, 88, 119

Di Rienzo, A., Peterson, A., Garza, J., Valdes, A., et al. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 91(8):3166, 1994. 10

Dohm, J., Lottaz, C., Borodina, T., and Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16):e105, 2008. 101

Drake, J., Charlesworth, B., Charlesworth, D., and Crow, J. Rates of spontaneous mutation. *Genetics*, 148(4):1667, 1998. 98

Durbin, R. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998. 57

Ellegren, H. Microsatellite mutations in the germline::: implications for evolutionary inference. *Trends in Genetics*, 16(12):551–558, 2000. 11

Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, 5(6):435–445, 2004. 11, 55

Ewing, B. and Green, P. Base-calling of automated sequencer traces usingPhred. II. error probabilities. *Genome research*, 8(3):186, 1998. 34

Ewing, B., Hillier, L., Wendl, M., and Green, P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome research*, 8(3):175, 1998. 36

Feuk, L., Carson, A., and Scherer, S. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006. 4

Gatchel, J. and Zoghbi, H. Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews Genetics*, 6(10):743–755, 2005. 11

Green, E., Guyer, M., et al. Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213, 2011. 2

Hamosh, A., Scott, A., Amberger, J., Bocchini, C., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514, 2005. 10

Hanawalt, P. Transcription-coupled repair and human disease. *Science*, 266(5193):1957, 1994. 116

Hazel, P., Huppert, J., Balasubramanian, S., and Neidle, S. Loop-length-dependent folding of G-quadruplexes. *Journal of the American Chemical Society*, 126(50):16405–16415, 2004. 115

Hoeijmakers, J. et al. Genome maintenance mechanisms for preventing cancer. *Nature*, 411(6835):366–374, 2001. 116

Homer, N., Merriman, B., and Nelson, S. BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, 4(11):e7767, 2009. 4

Hormozdiari, F., Alkan, C., Eichler, E., and Sahinalp, S. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome research*, 19(7):1270, 2009. 16

Huppert, J. and Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, 33(9):2908, 2005. 115

Huse, S., Huber, J., Morrison, H., Sogin, M., et al. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology*, 8(7):R143, 2007. 74

International Human Genome Sequencing Consortium, B. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001. 2

Kasai, K., Nakamura, Y., and White, R. Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its application to forensic science. *Journal of forensic sciences*, 35(5):1196, 1990. 87

Kashi, Y., King, D., and Soller, M. Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics*, 13(2):74–78, 1997. 11

Kashi, Y. and King, D. Simple sequence repeats as advantageous mutators in evolution. *TRENDS in Genetics*, 22(5):253–259, 2006. 11

Kelkar, Y., Tyekucheva, S., Chiaromonte, F., and Makova, K. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome research*, 18(1):30, 2008. 118

Koboldt, D., Chen, K., Wylie, T., Larson, D., et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283, 2009. 26

Korbel, J., Abyzov, A., Mu, X., Carriero, N., et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biology*, 10(2):R23, 2009. 15

Korbel, J., Urban, A., Affourtit, J., Godwin, B., et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420, 2007. 13, 32, 34

Kovtun, I. and McMurray, C. Features of trinucleotide repeat instability in vivo. *Cell research*, 18(1):198–213, 2008. 11

Krawitz, P., Rodelsperger, C., Jager, M., Jostins, L., et al. Microindel detection in short-read sequence data. *Bioinformatics*, 26(6):722, 2010. 4

Kuhn, R., Karolchik, D., Zweig, A., Trumbower, H., et al. The UCSC genome browser database: update 2007. *Nucleic acids research*, 35(suppl 1):D668, 2006. 29

Lai, Y. and Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution*, 20(12):2123, 2003. 87

Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nature methods*, 6(7):473–474, 2009. 82, 83

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome research*, 17(12):1787, 2007. 99

Lenzmeier, B. and Freudenreich, C. Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. *Cytogenetic and genome research*, 100(1-4):7–24, 2000. 11

Levy, S., Sutton, G., Ng, P., Feuk, L., et al. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, 2007. 2

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754, 2009. 4, 32

Li, H., Handsaker, B., Wysoker, A., Fennell, T., et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078, 2009. 73

Li, H., Ruan, J., and Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008. 4, 32

Lygo, J., Johnson, P., Holdaway, D., Woodroffe, S., et al. The validation of short tandem repeat (STR) loci for use in forensic casework. *International Journal of Legal Medicine*, 107(2):77–89, 1994. 87

Madsen, B., Villesen, P., and Wiuf, C. Short tandem repeats in human exons: a target for disease mutations. *BMC genomics*, 9(1):410, 2008. 10

Mahtani, M. and Willard, H. A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci. *Human molecular genetics*, 2(4):431, 1993. 11

Mardis, E. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008. 13

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297, 2010. 4

McKernan, K., Peckham, H., Costa, G., McLaughlin, S., et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19(9):1527, 2009. 4

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D., et al. Identification and correction of systematic error in high-throughput sequence data. 2011. 101

Medvedev, P., Stanciu, M., and Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *nature methods*, 6:S13–S20, 2009. 5

Metzker, M. Sequencing technologiesthe next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009. 3

Mills, R., Luttig, C., Larkins, C., Beauchamp, A., et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, 16(9):1182, 2006. 25

Moore, G. et al. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998. 1

Myers, E. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79, 2005. 28

Nachman, M. and Crowell, S. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297, 2000. 98

Nature Jobs. New opportunities in the genomic era. 2011. 2

Ning, Z., Cox, A., and Mullikin, J. SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10):1725, 2001. 36

Ning, Z., Spooner, W., Spargo, A., Leonard, S., et al. The SSAHA trace server. 2004. 36

Pearson, C., Edamura, K., and Cleary, J. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10):729–742, 2005. 10, 11

Pearson, W. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11(3):635–650, 1991. 36

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org`. ISBN 3-900051-07-0. 104

Ruitberg, C., Reeder, D., and Butler, J. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Research*, 29(1):320, 2001. 87

Rumble, S., Lacroute, P., Dalca, A., Fiume, M., et al. SHRiMP: accurate mapping of short color-space reads. *PLoS computational biology*, 5(5):e1000386, 2009. 4

Sanger, F., Coulson, A., Hong, G., Hill, D., et al. Nucleotide sequence of bacteriophage [lambda] DNA. *Journal of Molecular Biology*, 162(4):729–773, 1982. 3

Sindi, S., Helman, E., Bashir, A., and Raphael, B. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):i222, 2009. 15

Sprecher, C., Puers, C., Lins, A., and Schumm, J. General approach to analysis of polymorphic short tandem repeat loci. *BioTechniques*, 20(2):266–277, 1996. 88

The Economist. Data, data everywhere. 2010. 1

Tuzun, E., Sharp, A., Bailey, J., Kaul, R., et al. Fine-scale structural variation of the human genome. *Nature genetics*, 37(7):727–732, 2005. 13, 15

Urquhart, A., Kimpton, C., Downes, T., and Gill, P. Variation in short tandem repeat sequencesa survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine*, 107(1):13–20, 1994. 87

Volik, S., Raphael, B., Huang, G., Stratton, M., et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome research*, 16(3):394, 2006. 13

Wang, J., Wang, W., Li, R., Li, Y., et al. The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, 2008. 4

Weber, J. and Wong, C. Mutation of human short tandem repeats. *Human molecular genetics*, 2(8):1123, 1993. 87

Wheeler, D., Srinivasan, M., Egholm, M., Shen, Y., et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 2008. 4

Whittaker, J., Harbord, R., Boxall, N., Mackay, I., et al. Likelihood-based estimation of microsatellite mutation rates. *Genetics*, 164(2):781, 2003. 87

Xu, H., Chakraborty, R., and Fu, Y. Mutation rate variation at human dinucleotide microsatellites. *Genetics*, 170(1):305, 2005. 88

Zeitouni, B., Boeva, V., Janoueix-Lerosey, I., Loeillet, S., et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, 26(15):1895, 2010. 16

Zhivotovsky, L., Rosenberg, N., and Feldman, M. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *The American Journal of Human Genetics*, 72(5):1171–1186, 2003. 11