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).

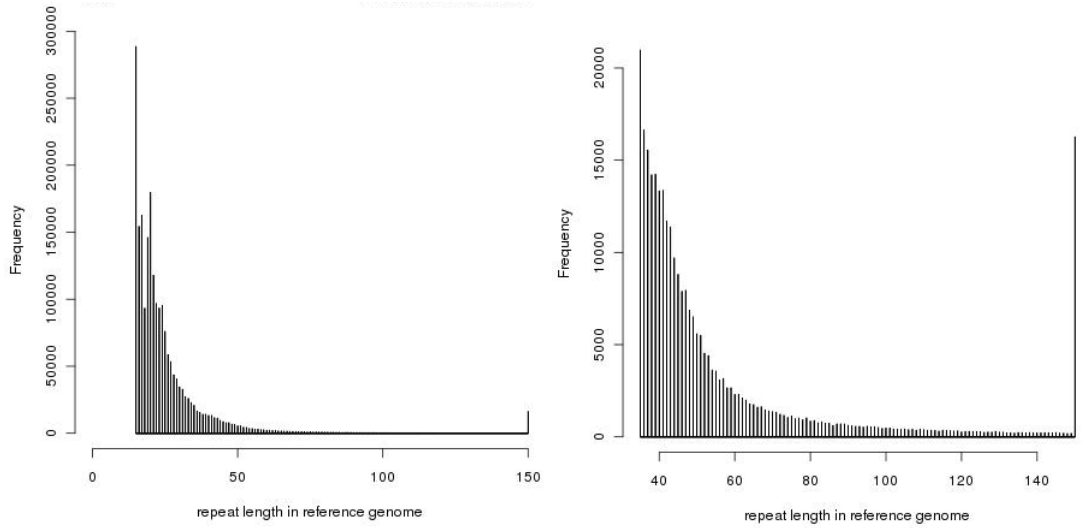
The detection step is broken down into a series of algorithms which scan through the genome looking for repetitive patterns known as $k-tuples$. A $k-tuple$ is a window of k consecutive characters from a nucleotide sequence. Matching $k-tuples$ are two windows with identical contents and if aligned in 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.

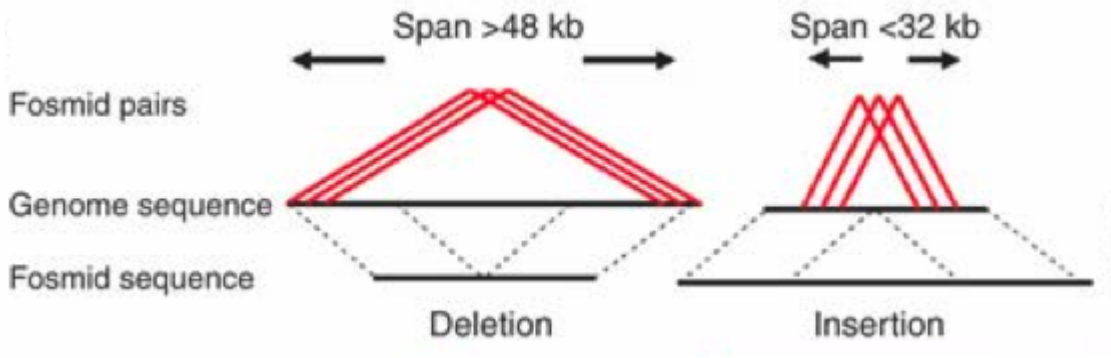


Figure 1.2: The underlying paired end sequencing methodology used to detect structural variation by fosmid pairing (Tuzun et al. [2005]). Deletions in the fosmid source are defined as sites where two or more fosmid end-sequence pairs span $>$ 48 kb. Insertions are defined as sites where two or more fosmids 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

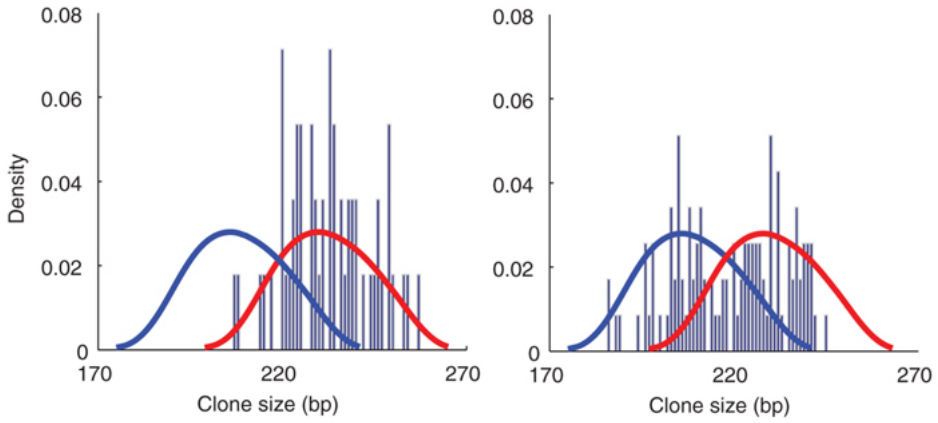
and statistics that defines a sequence – or other collection of random variables – that each random variable has the same probability distribution as the others and are mutually independent. From this, we are able to define the K-S statistic for a given cumulative distribution function $F(x)$ as

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

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.

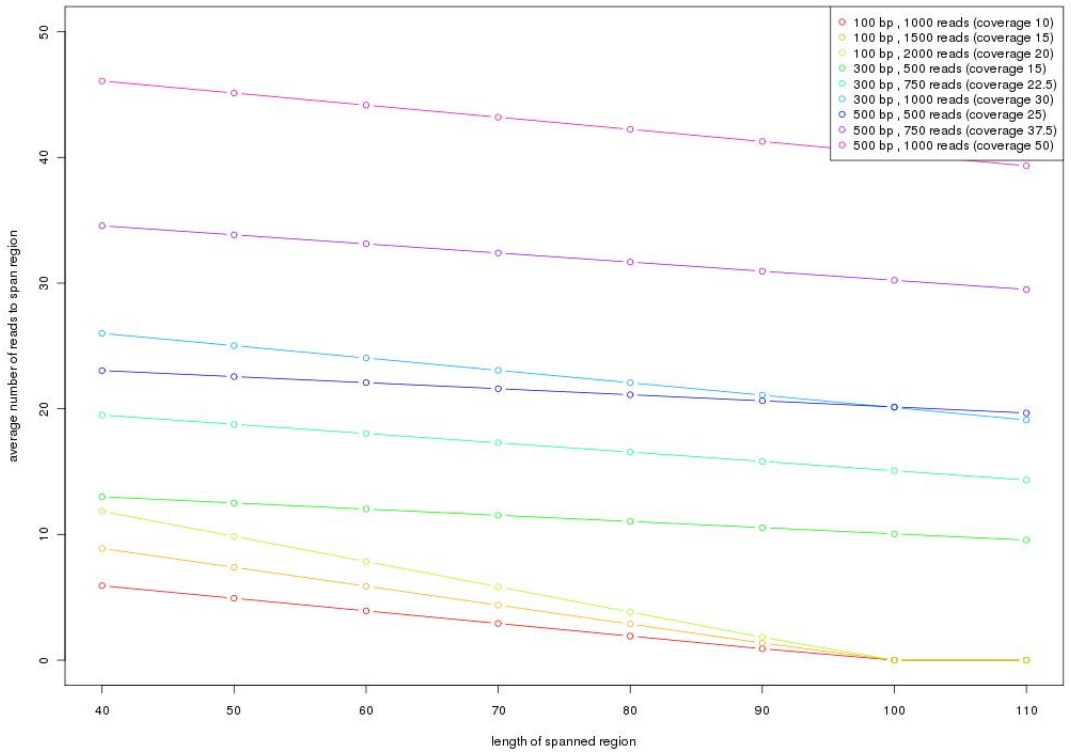


Figure 1.4: Graph of expected number of spanning reads (physical coverage) and reads that extend across various genomic lengths at base pair coverages of 10, 15 and 20x. Reads of length of 100 bp were chosen to illustrate the upper bound of read lengths currently available. The spanning coverage was 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

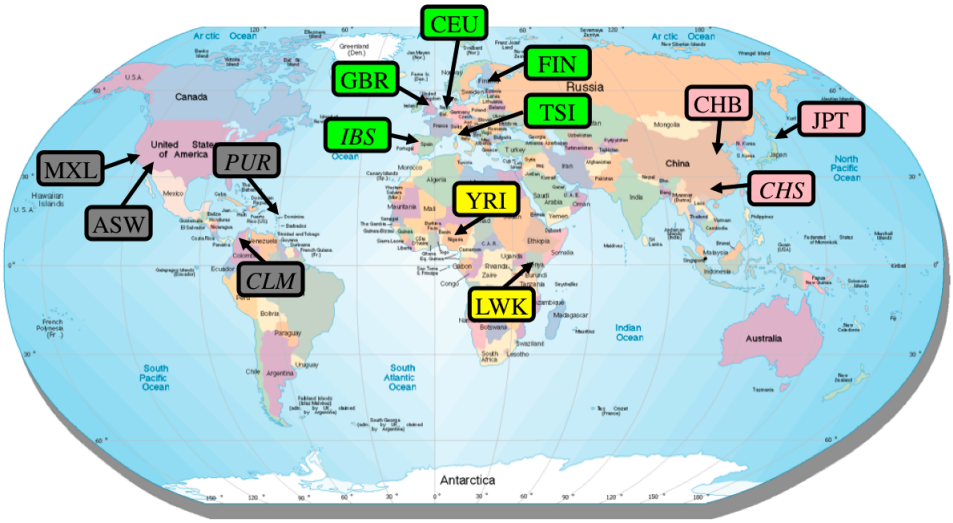


Figure 1.5: Map of populations in 1000 Genomes Project Phase 1 build.

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.