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

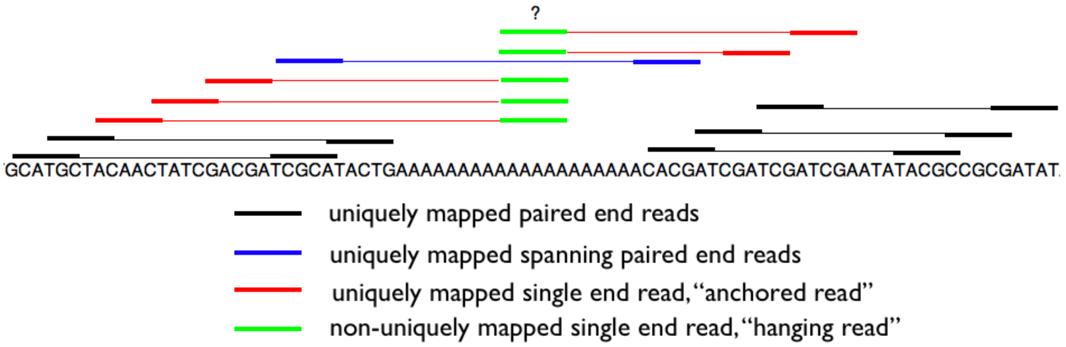

?
GCATGCTACAACTATCGACGATCGCATACTGAAAAAAAAAAAAAAAAAAAACACGATCGATCGATCGAATATACGCCGCGATAT
uniquely mapped paired end reads
uniquely mapped spanning paired end reads
uniquely mapped single end read, "anchored read"
non-uniquely mapped single end read, "hanging read"

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 ≤ 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 (∼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 $\geq 97\%$ identity to both flanking regions and the matches were in the same order as in the reference genome; (ii) the matches covered $\geq 80\%$ 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(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

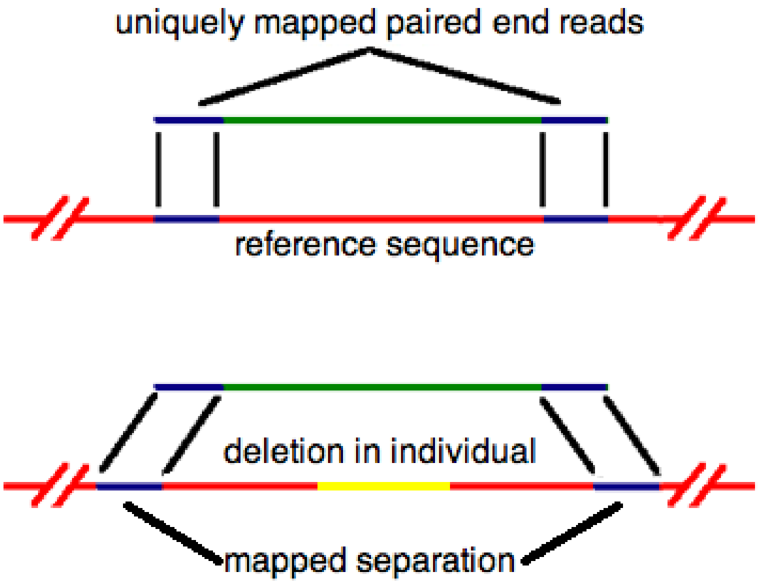


Figure 2.2: Graphic of mapped paired end read alignments of an individual whose locus matches the reference (top) and whose locus contains a deletion in respect to the reference (bottom). The blue paired end reads at top align at the exact distance one would expect to observe given the fragment length 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-

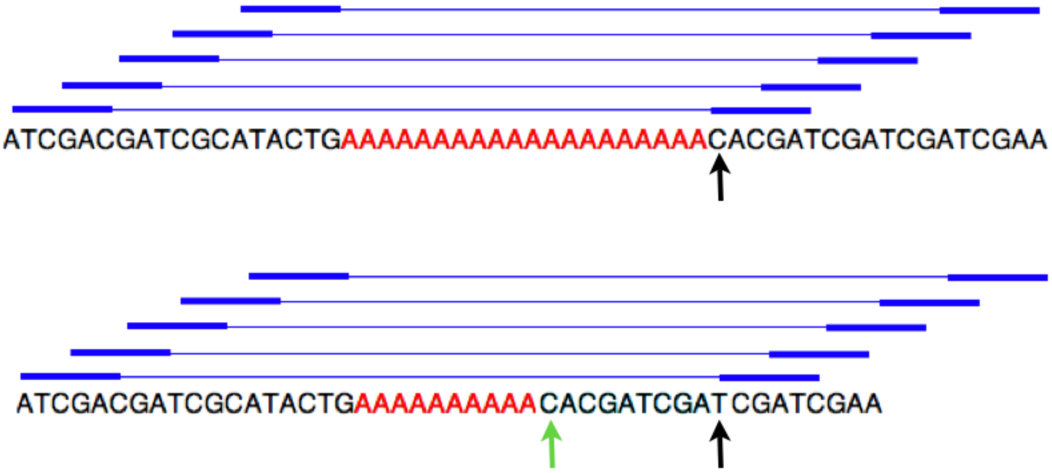
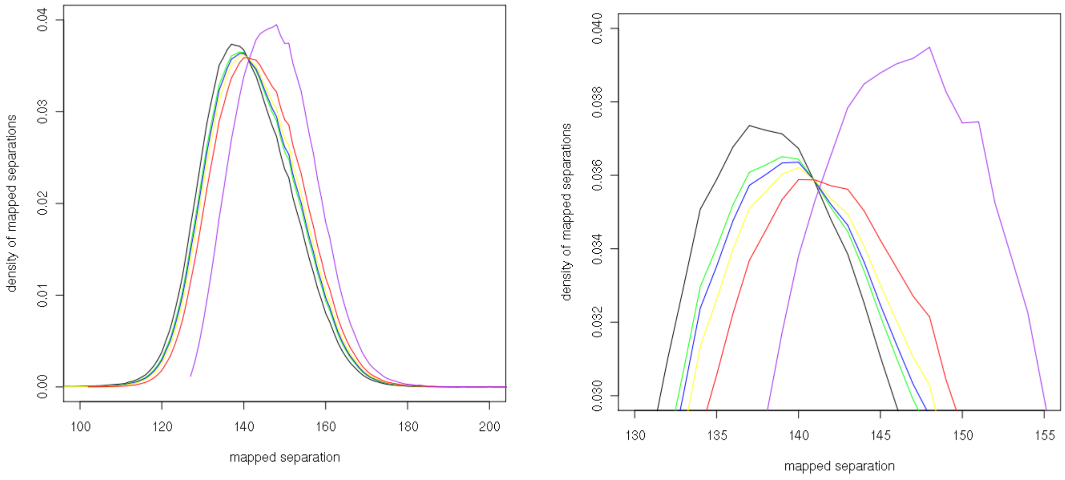

ATCGACGATCGCATACTGAAAAAAAAAAAAAAAAAAAACACGATCGATCGATCGAA
ATCGACGATCGCATACTGAAAAAAAAAACACGATCGATCGATCGAA

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

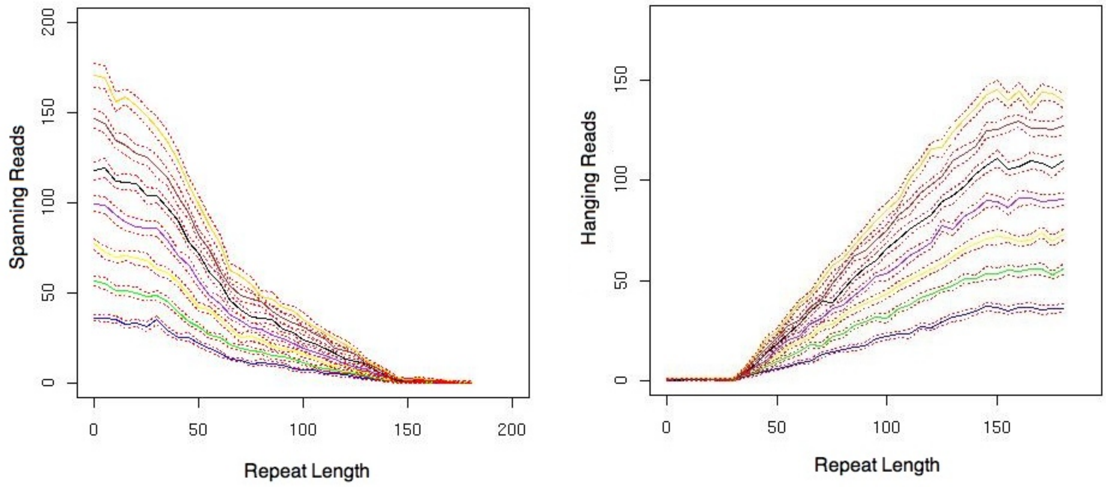


(a) Distribution of MPERS for library g1k-sc-NA12878-CEU-1 across differing repeat lengths

(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.

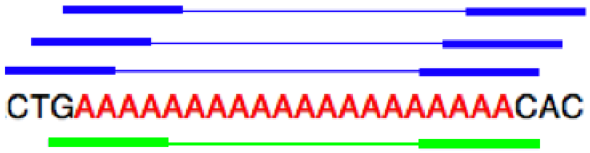


Figure 2.6: Cartoon representation of actual mapping positions of two paired end reads across a poly-A repeat of length 20 bp. The blue paired end read of MPERS 24 bp has three unique positions it can map to and still uniquely span the repeat, while the green paired end read of MPERS of 22 bp has only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

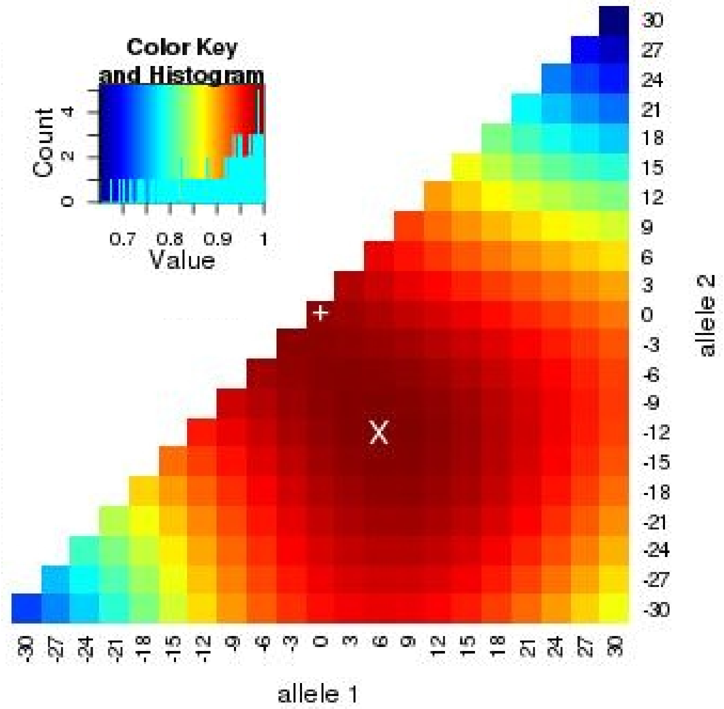


Figure 2.7: Heatmap of likelihoods at a selected repeat locus of length 60 bp from a simulated homozygous reference genotype with average base pair coverage 15x. In the simulation, the maximum likelihood estimate determined the genotype to be {6,-12} (white X), where the actual genotype was reference (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.

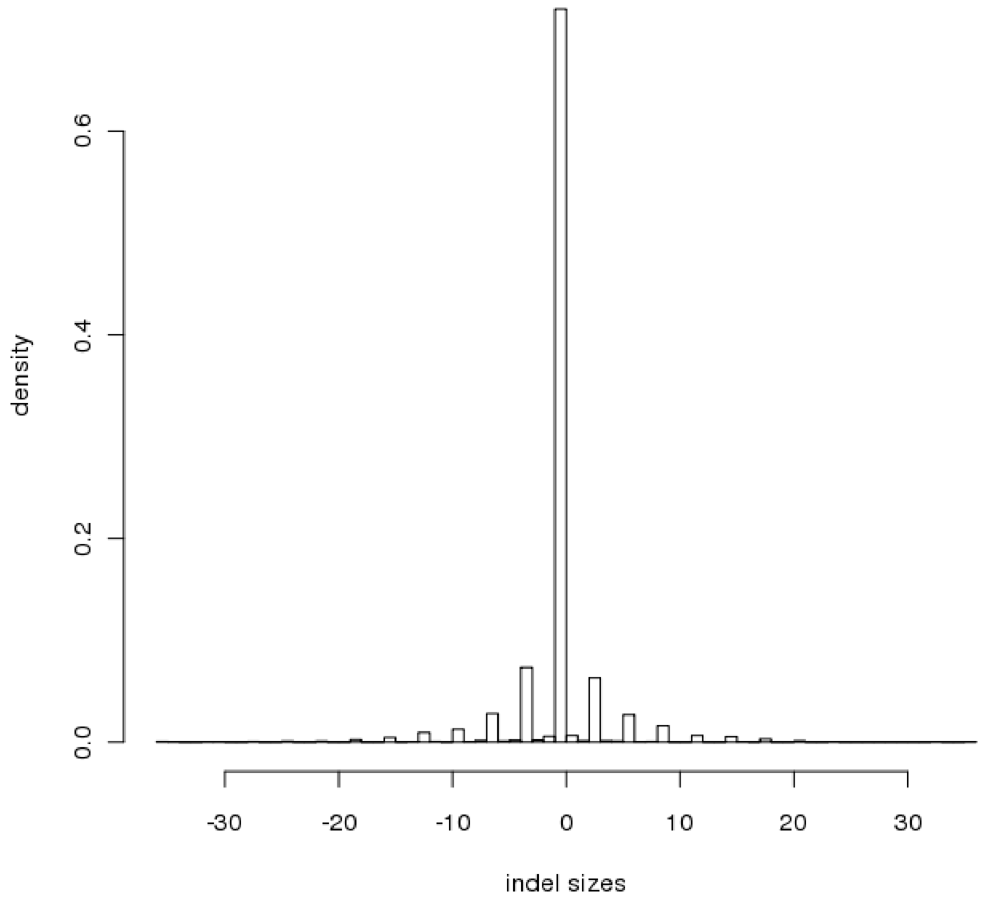


Figure 2.8: Prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads. This distribution is based on 3,435 calls; 2,474 reference calls (72.0%) and 961 indel calls (28.0%).

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

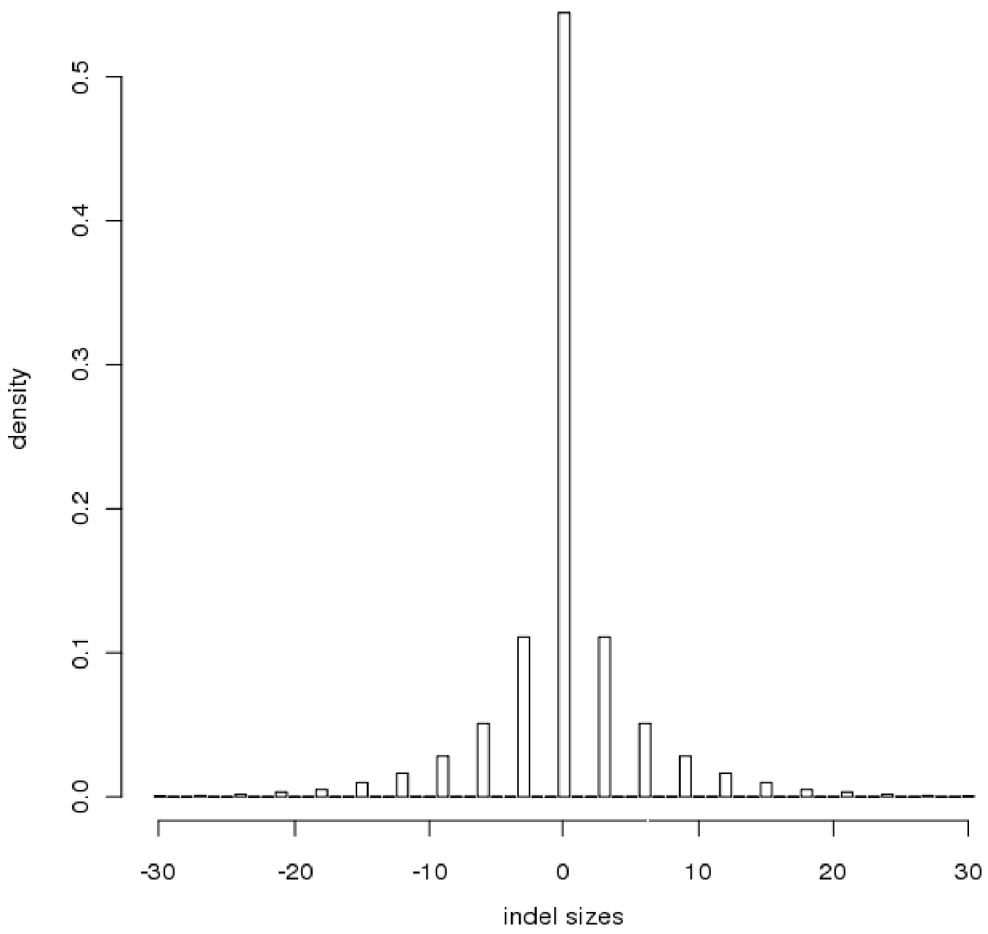


Figure 2.9: Symmetric prior probability distribution of haploid indel calls made in individual NA18507 from capillary reads.

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 $k = 3$.

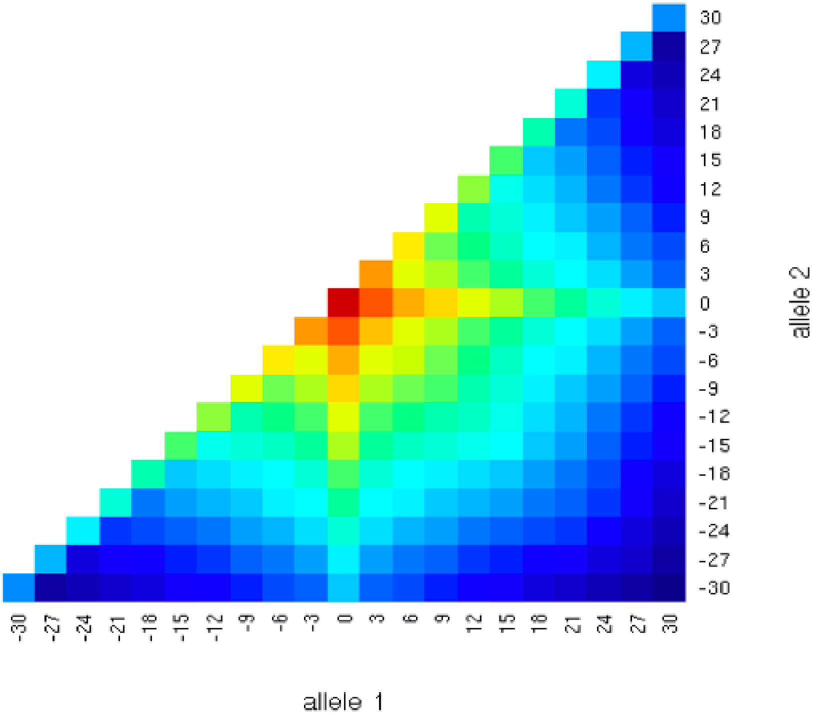


Figure 2.10: Heat map of the estimated prior probabilities for the varying genotypes of a triplet repeat locus in an individual (shown in log space). The confirmation of the probability space illustrates the assumptions made about which genotypes will be more likely than others. The most probable genotype 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).

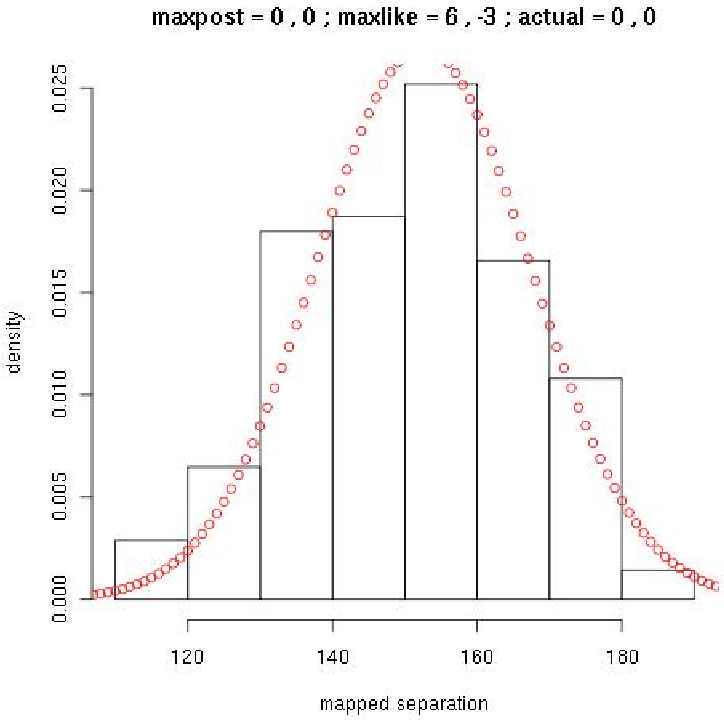


Figure 2.11: Histogram of spanning paired end read separations and MPERS distribution for a reference genotype simulation. The histogram of spanning paired end read separations across a simulated repeat tract coincide distinctly with the distribution of MPERS you would expect to observe given the sample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

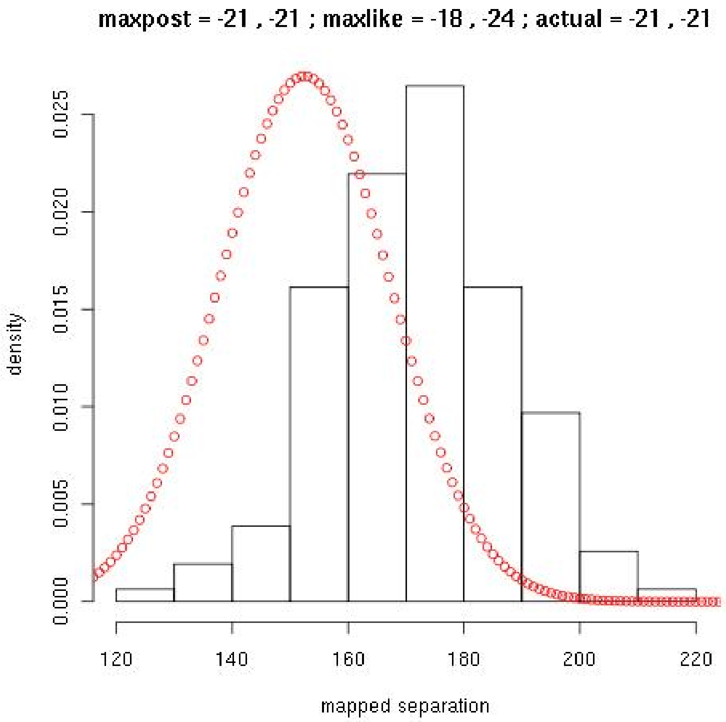


Figure 2.12: Histogram of spanning paired end read separations across an individual whose repeat length is 21 bp shorter than that in the reference graphed against the MPERS distribution for a reference length genotype. The mean MPERS for the spanning paired end reads is therefore shifted approximately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

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.

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

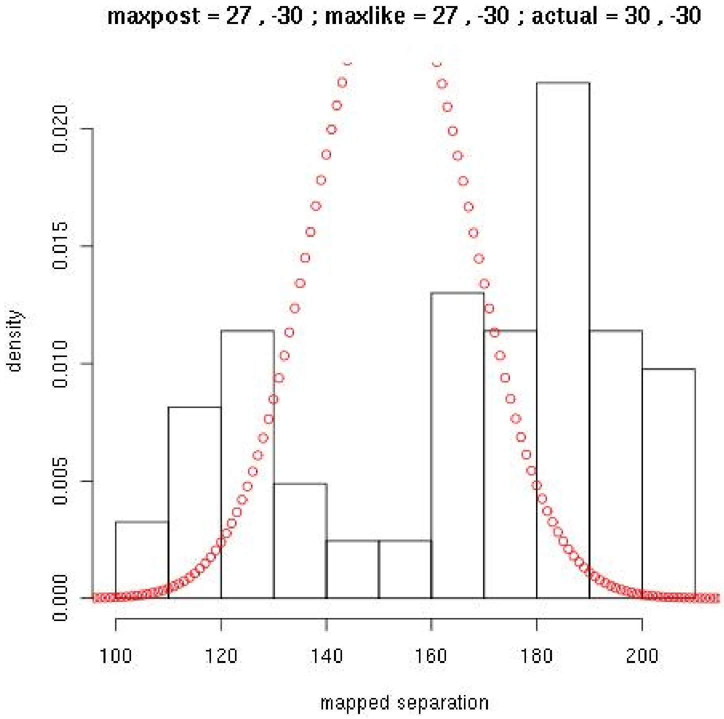


Figure 2.13: Histogram of spanning paired end reads across an individual whose two copies differ in length from the reference graphed against the MPERS distribution for a reference length genotype. This sample contains an insertion of 30 bp and a deletion of 30 bp. It is quite obvious that the number 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 chapter 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.

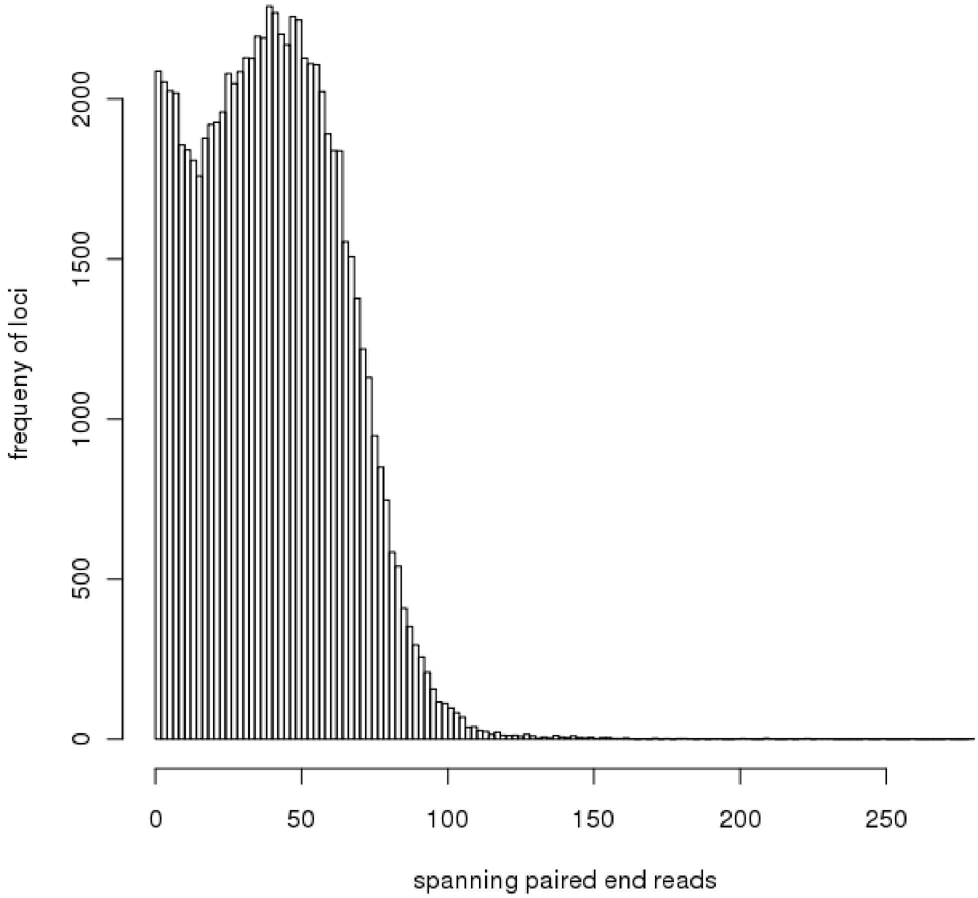


Figure 2.14: Histogram for loci containing a given number of spanning paired end reads for every triplet repeat loci in individual NA12878. The smaller libraries' number of spanning paired end reads will diminish much quicker than the larger fragment libraries and this could explain why the histogram 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

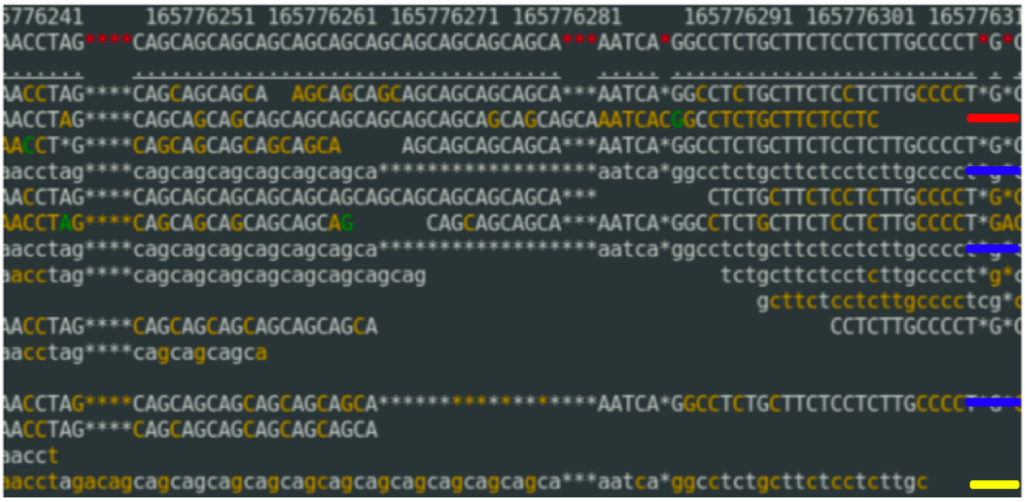
As part of the pilot project for the 1000 Genomes Project, individual NA12878 was sequenced to approximately 22.5x depth using the Illumina sequencer (Consortium [2010]). The same DNA was also sequenced using the 454 sequencer and capillary sequence to approximately 12.8x and 0.5x depth, respectively. 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	111
		homozygous indel	0
		heterozygous	2
		inconclusive	1
homozygous indel	158	reference	5
		homozygous indel	56
		heterozygous	52
		inconclusive	45
heterozygous	64	reference	4
		homozygous indel	3
		heterozygous	44
		inconclusive	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]).

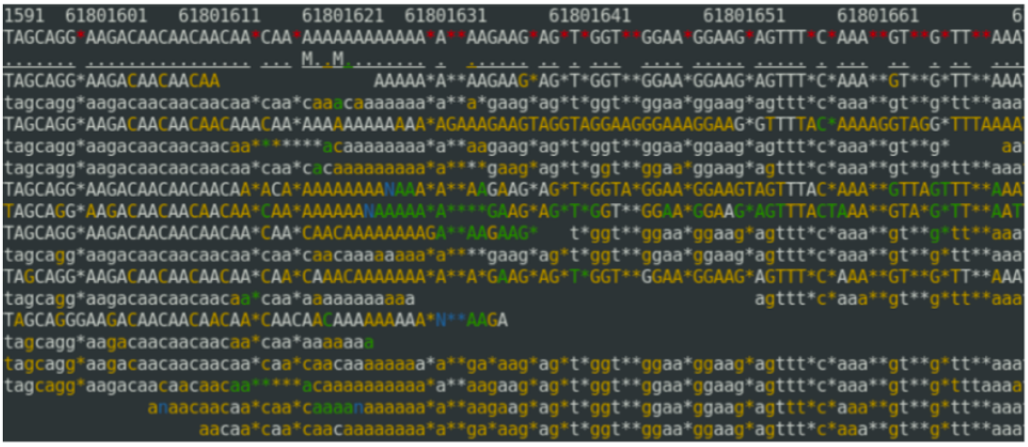
After visual inspection of the 454 alignments in tview, we removed any sites where the alignments remained unclear (59 sites in total were removed, table 2.5). Figures 2.16 and 2.15 depict two loci where we are both able and unable, respectively, to make the correct genotype call based on visual inspection.



```
5776241     165776251 165776261 165776271 165776281     165776291 165776301 16577631
AACCTAG****CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA***AATCA*GGCCTCTGCTTCTCCTCTTGCCCCT*G*C
.......    ...................................   ..... ......................... . .
AACCTAG****CAGCAGCAGCA  AGCAGCAGCAGCAGCAGCAGCA***AATCA*GGCCTCTGCTTCTCCTCTTGCCCCT*G*C
AACCTAG****CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAAATCACGGCCTCTGCTTCTCCTC
AACCT*G****CAGCAGCAGCAGCAGCA     AGCAGCAGCAGCA***AATCA*GGCCTCTGCTTCTCCTCTTGCCCCT*G*C
aacctag****cagcagcagcagcagcagca******************aatca*ggcctctgcttctcctcttgcccc
AACCTAG****CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA***         CTCTGCTTCTCCTCTTGCCCCT*G*C
AACCTAG****CAGCAGCAGCAGCAGCAG      CAGCAGCAGCA***AATCA*GGCCTCTGCTTCTCCTCTTGCCCCT*GAC
aacctag****cagcagcagcagcagcagca******************aatca*ggcctctgcttctcctcttgcccc
aacctag****cagcagcagcagcagcagcagcag                        tctgcttctcctcttgcccct*g*c
                                                              gcttctcctcttgcccctcg*c
AACCTAG****CAGCAGCAGCAGCAGCAGCA                                     CCTCTTGCCCCT*G*C
aacctag****cagcagcagca

AACCTAG****CAGCAGCAGCAGCAGCAGCA******************AATCA*GGCCTCTGCTTCTCCTCTTGCCCC
AACCTAG****CAGCAGCAGCAGCAGCAGCA
aacct
aacctagacagcagcagcagcagcagcagcagcagcagcagcagca***aatca*ggcctctgcttctcctcttgc
```

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).



```
1591  61801601   61801611    61801621  61801631      61801641        61801651      61801661        6
TAGCAGG*AAGACAACAACAACAA*CAA*AAAAAAAAAAAA*A**AAGAAG*AG*T*GGT**GGAA*GGAAG*AGTTT*C*AAA**GT**G*TT**AAA
........ .................. ... M..M......... .  ....... .. . ...  .... ...... ...... . ... .. . .. ...
TAGCAGG*AAGACAACAACAA              AAAAA*A**AAGAAG*AG*T*GGT**GGAA*GGAAG*AGTTT*C*AAA**GT**G*TT**AAA
tagcagg*aagacaacaacaacaa*caa*caaacaaaaaaa*a**a*gaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
TAGCAGG*AAGACAACAACAACAAACAA*AAAAAAAAAAAA*AGAAAGAAGTAGGTAGGAAGGGAAAGGAAG*GTTTTAC*AAAAGGTAGG*TTTAAAA
tagcagg*aagacaacaacaacaa*******acaaaaaaaa*a**aagaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*     aa
tagcagg*aagacaacaacaacaa*caa*cacaaaaaaaaa*a****gaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
TAGCAGG*AAGACAACAACAACAA*ACA*AAAAAAAANAAA*A**AAGAAG*AG*T*GGTA*GGAA*GGAAGTAGTTTAC*AAA**GTTAGTTT**AAA
TAGCAGG*AAGACAACAACAACAA*CAA*AAAAAANAAAAA*A****GAAG*AG*T*GGT**GGAA*GGAAG*AGTTTACTAAA**GTA*G*TT**AAT
TAGCAGG*AAGACAACAACAACAA*CAA*CAACAAAAAAAAGA**AAGAAG*   t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
tagcagg*aagacaacaacaacaa*caa*caacaaaaaaaa*a****gaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
TAGCAGG*AAGACAACAACAACAA*CAA*CAAACAAAAAAA*A**A*GAAG*AG*T*GGT**GGAA*GGAAG*AGTTT*C*AAA**GT**G*TT**AAA
tagcagg*aagacaacaacaacaa*caa*aaaaaaaaaa                                  agttt*c*aaa**gt**g*tt**aaa
TAGCAGGGAAGACAACAACAACAA*CAACAACAAAAAAAAA*N**AAGA
tagcagg*aagacaacaacaacaa*caa*aaaaaaa
tagcagg*aagacaacaacaacaa*caa*caacaaaaaaaa*a**ga*aag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
tagcagg*aagacaacaacaacaa*****acaaaaaaaaaa*a**aagaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tttaaaa
                 anaacaacaa*caa*caaaanaaaaaa*a**aagaag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
                      aacaa*caa*caacaaaaaaaaa*a**ga*aag*ag*t*ggt**ggaa*ggaag*agttt*c*aaa**gt**g*tt**aaa
```

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).

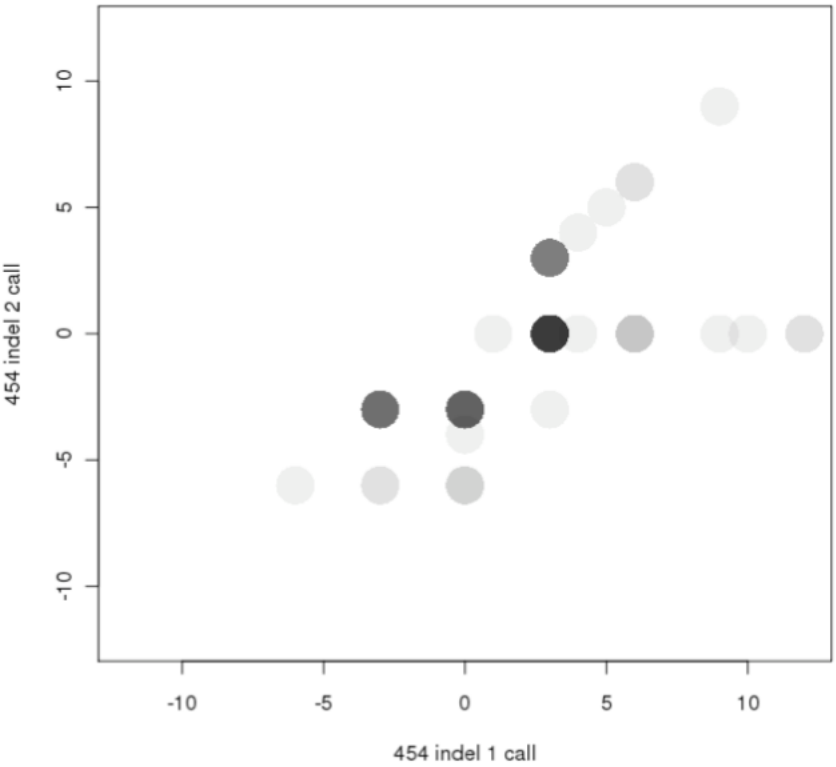


Figure 2.17: Plot of the 454 indel genotypes when our method called a reference genotype, {0, 0}. Almost all these genotypes' repeat lengths are within ±3 bp of the reference length (86%) as demarcated by the four dark black dots around the reference call. As the absolute difference between the reference 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

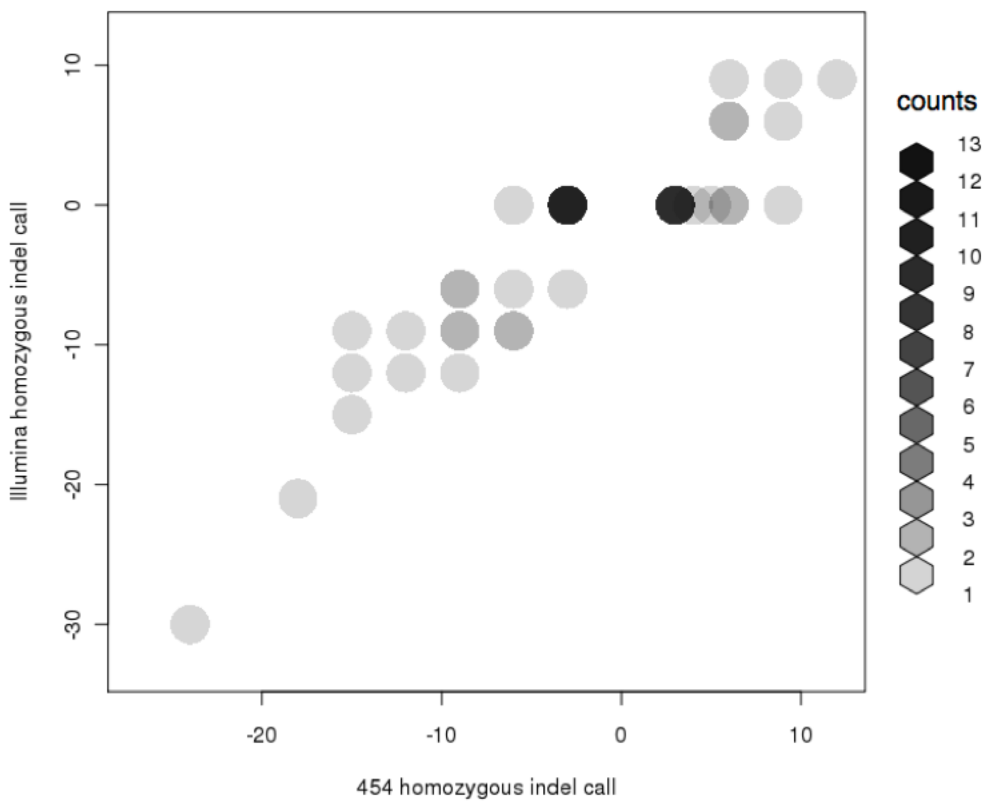


Figure 2.18: Comparison of true homozygous indel genotypes as called from 454 sequence to that of our method's calls at these loci. The diagonal $x = y$ represents our methods calls being exactly on the true genotype call with any deviations from this line an error. Most all calls are within 6 bp from 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.

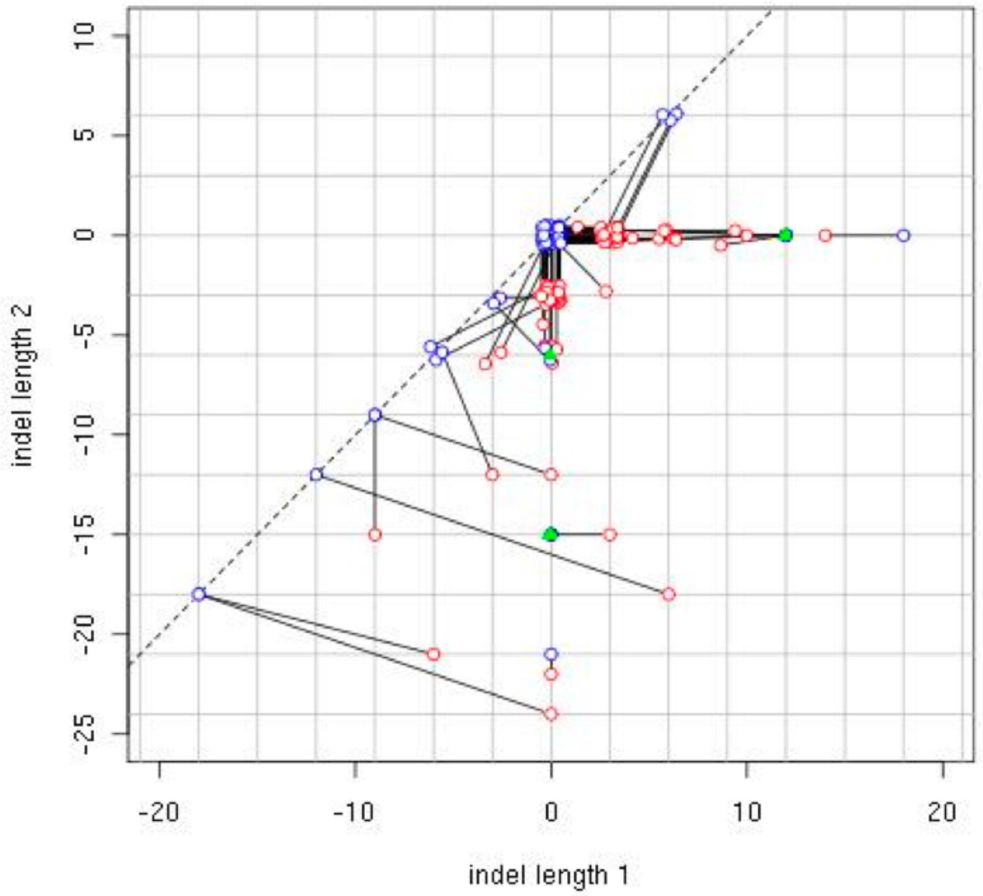


Figure 2.19: Join plot comparison of actual genotype (red dot) compared to the genotype called by our method (blue dot). The dotted diagonal line represents the homozygous genotype while the solid black lines illustrate the difference in genotype calls between the truth and out method's call. Ideally, 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 $x_1 = y_1 = 0$

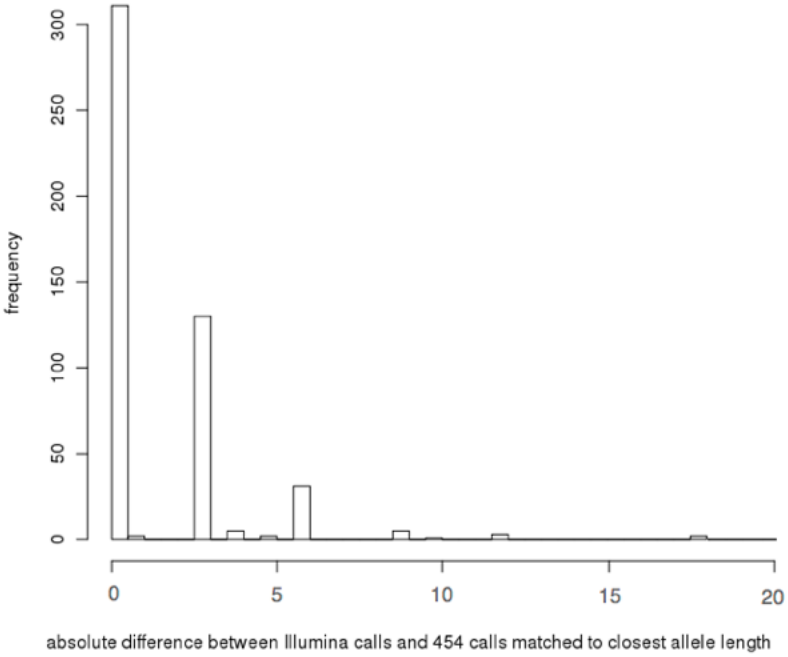


Figure 2.20: Histogram of differences in proximal allele lengths between genotype calls made by 454 and our method. More than half (63%) of all of our method's allele lengths match exactly the allele length inferred from 454 reads. When the threshold is raised to ±3 bp, the percentage raises further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

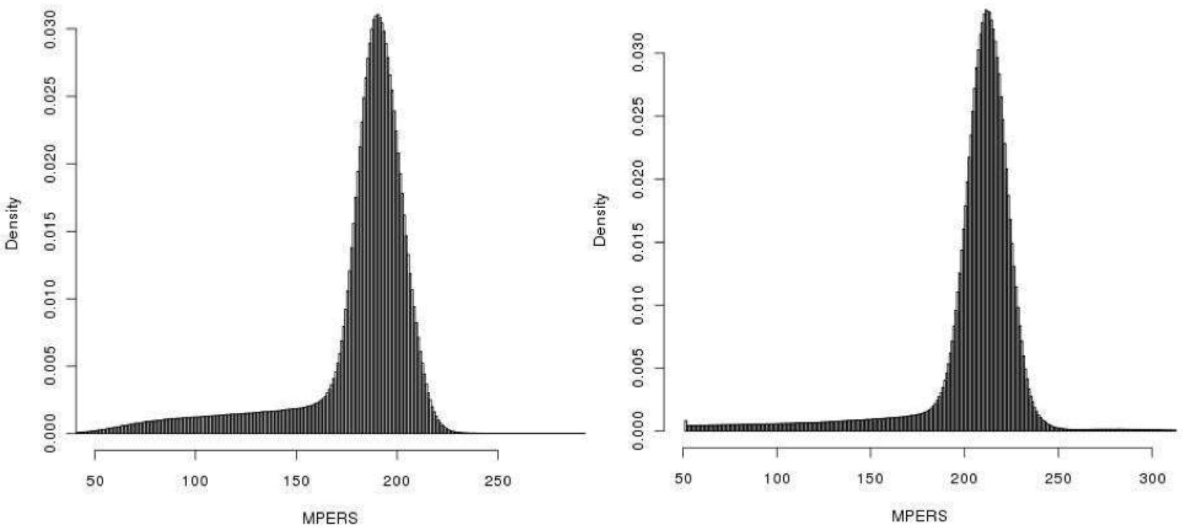
We have developed a novel method which uses short paired end read sequencing data to infer the genotype (repeat length) of the two copies of a STR locus in a diploid individual. Our method estimates the lengths of the two indels – one in each of the two copies of the individuals locus – and then calculates 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 ≥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 ±3 bp (see figure 2.20). One limitation of our method is that it requires a reasonable number of spanning paired end reads (≥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.