Chapter 3

Factors influencing polymorphism in short tandem repeats

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

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

Past research has sought to understand their evolution over time (Calafell et al. [1998]) as well as use STRs as markers for forensic analysis (Kasai et al. [1990]; Urquhart et al. [1994]; Lygo et al. [1994]; Ruitberg et al. [2001]). As of the writing of this work, there has been no genome wide assay of short tandem repeats that we are aware of. A genome wide assay of STRs would have the power to elucidate what factors in STRs increase the chance of observing a variant at a locus. Some of the proposed factors include the composition of the repeat motif, the purity of the repeat in the reference genome, the length of the repeat in the reference genome, the GC content of the repeat and proximal sequence and whether the STR resides within a transcript. There has been past research that looked into understanding how some of these factors affect mutation rate (Xu et al. [2005]), but nothing on a large, genome wide scale. This is due mostly in part to the fact that past sequencing of STRs is both costly and slow (Sprecher et al. [1996]), which has precluded a large, genome wide assay. However, due to the advent of next generation sequencing technology, we are now able to explore these loci on a massive scale.

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

3.1 Sources of sequence

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

3.1.1 1000 Genomes pilot trios

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

3.1.1.1 Sequencing statistics

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

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

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

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

3.1.2 Illumina Trio

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

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

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

3.2 MPERS distributions

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

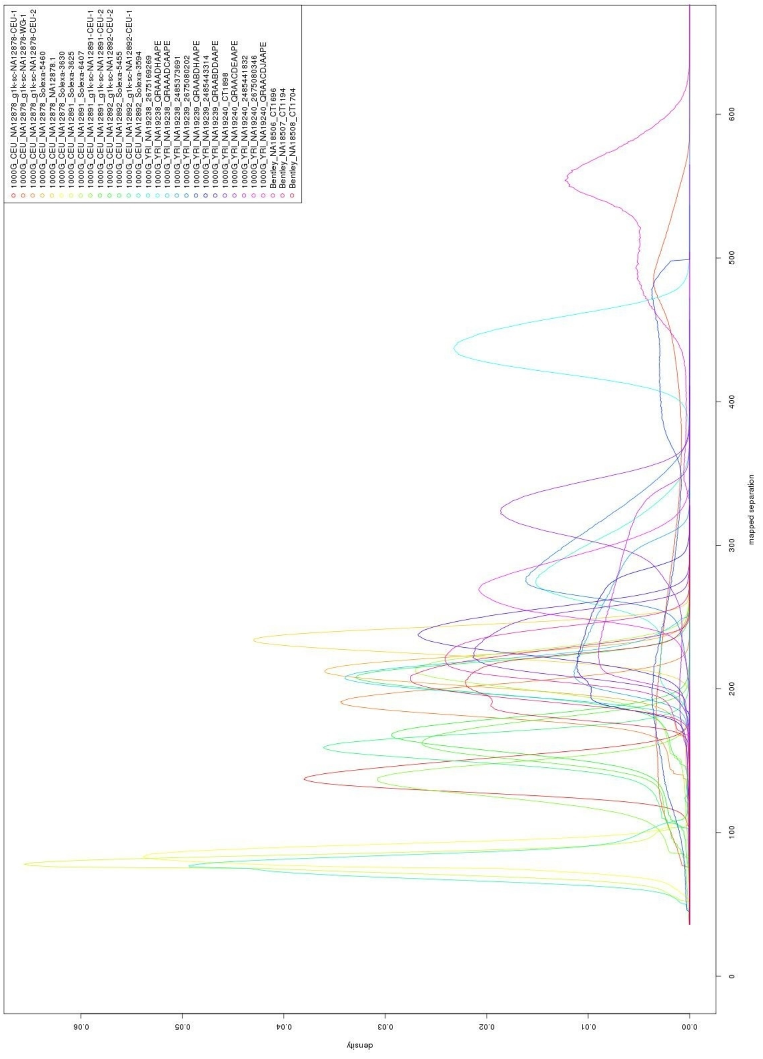

Figure 3.1: Distributions of each library in the nine individuals from the three trios data set. Made up of thirty libraries, the range and shape of each library is unique. The libraries which yield more information for our analysis are those that are tightly distributed around the fragment size library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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3 Detecting indels in short tandem repeats

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

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

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

3.4 Short tandem repeat criteria

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

3.4.1 STR metrics

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

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

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

3.4.2 Tandem repeat length in reference (reflen)

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

$$l = z - y + 1$$

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

3.4.3 Tandem repeat motif family (motif)

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

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

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

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

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

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

3.4.4 Purity of tandem repeat in reference

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

3.4.4.1 Longest pure stretch (purls)

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

3.4.4.2 Percent match (purnew)

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

For our analysis, we devised two related metrics to measure the percent match a tandem repeat had to its given repeat motif. The first, purnew, is the overall adherence of a STRs sequence to its repeat motif. This algorithm looks at each subsequence of length of the motif and determines if it matches the overall consensus motif pattern. The algorithm calculates the proportion of start positions in a tandem repeat locus whose subsequent sequence matches the family of motifs a repeat locus is attributed to. It should be noted, however, that it only gives a positive score for subsequences that match the motif on the same strand. For instance, the family of motifs AAC would have AAC, ACA and CAA on the forward strand and TTG, GTT and TGT on the reverse. If the motifs of the reverse strand appear on the forward strand, they are considered foreign bases and not scored as fitting the motif pattern. The second metric, lenpurnew, is simply the value of purnew multiplied by the length of the repeat locus in the reference. This in essence scales the percent match value to the repeat length. We believed it was important to have this additional metric associated with the purnew metric because ignoring the length of the STR gives rise to a bias in shorter STRs having a higher purity metric score than longer STRs. This bias is described later in section 3.5.1.1.

3.4.4.2.1 Percent match algorithm

```
1. Set $score = 0$
2. Define all possible permutations which match a family of motifs that reside
   on the same strand
3. Starting at $x = 1$
4. If subsequence $(S_x, ..., S_{x+|M|-1})$ matches a possible permutation defined in
   2, $score++$
5. $x++$
       If $x \leq |S| - |m| + 1$
         goto 4
       else
         last
6. Calculate purity as $\frac{score}{|S|-|m|+1}$
```

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

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

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

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

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

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

3.5 Results

We approached our analysis of STR factors in two ways: the influence each factor had on observing a non-reference allele, and the effect each factor had on the overall magnitude of the observed indel for both insertions and deletions. To begin, we sought to determine the effects of observing a non-reference allele by using a logistic regression which determined the influence each factor had on observing a non-reference allele at a given locus. In more detail, a logistic regression is used in predicting the probability of the occurrence of an event by fitting the data to a logit function of a logistic curve. For our purposes, we were interested in the logistic regression as it is a generalized linear model (GLM) used in binomial regressions (discussed below). Like other regressions, the logistic linear regression can make use of several predictor variables (our factors) that may be either numerical (purity, reference length, etc.) or categorical (motif family, trio family, etc.).

To begin, the logistic function is defined as

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

where z is some linear relationship between the explanatory variables

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

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

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

where D represents the response data, $\mathbb{I}_{y_i}$ is the indicator function which takes the value one when an event occurs and zero otherwise. The likelihood function is specified by defining the parameters μ_i as functions of the explanatory variables (in our case the factors). There are many methods of generating the values of μ in systematic ways that allow for interpretation of the model. However, there is a requirement that the model linking the probabilities μ to the explanatory variables should be of a form which only produces values in the range 0 to 1 which we have described above in equation 3.2. It is then only a matter of fitting the model to the parameter values that maximize the likelihood in equation 3.3.

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

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

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

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

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

3.5.1 Modeling of factors

A logistic regression was used to determine the effect each of the 11 factors had on observing a non-reference allele in a STR locus. For ease of computation and modeling, we separated the called genotypes into two alleles and did all the analysis at the level of alleles. This appeared to be the easiest approach and we did not feel it changed the overall inference we could make regarding the the outcome of our modeling. A summary of the model's output is produced by R, giving the value of each of the coefficients for each of the factors as well as a p-value that indicated the confidence the model had that each of the coefficient values was non-zero. For almost every coefficient calculated in our analysis, the p-value was less than 0.001. Because of this, when we discuss specific coefficients below they will by default have a p-value less than 0.001. In the rare cases where this isn't the case, we shall explicitly state which factors' coefficients are not statistically significant. This is the same for our linear model which we used to determine what effect, if any, a factor has on the magnitude of an observed indel.

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

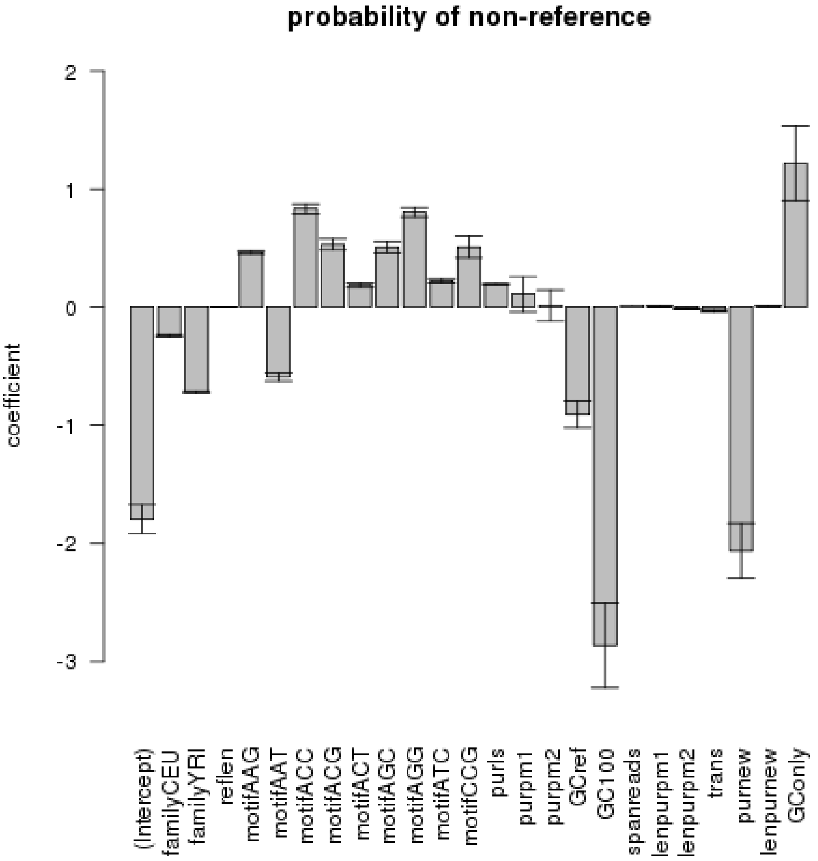

Figure 3.2: Graph of coefficients determined by full logistic regression of factors giving contradictory results because of confounding between correlated factors.

By sorting the scaled coefficients by the absolute value and plotting them on the same graph, it was clear which factors had the largest effect (be it positive or negative). In the end, we ended up with four plots: logistic regression for non-reference, linear regression for the magnitude of an indel and linear regressions for the size of both insertions and deletions. The graphs of each of these scenarios are plotted in figures 3.3, 3.4, 3.5 and 3.6 which illustrate the absolute effect of each of the factors. On each graph, all the coefficient values are shown aside from those having a p-value > 0.05 which include motifs ACG and AGC in the logistic linear model, motifs ACG and ACT in the insertions linear model and trans in the deletions linear model. Out of all the factors' coefficients that were graphed, all had a p-value < 0.001 except for GCref in the insertions linear model that had a p-value in between the range of $(0.01, 0.05)$.

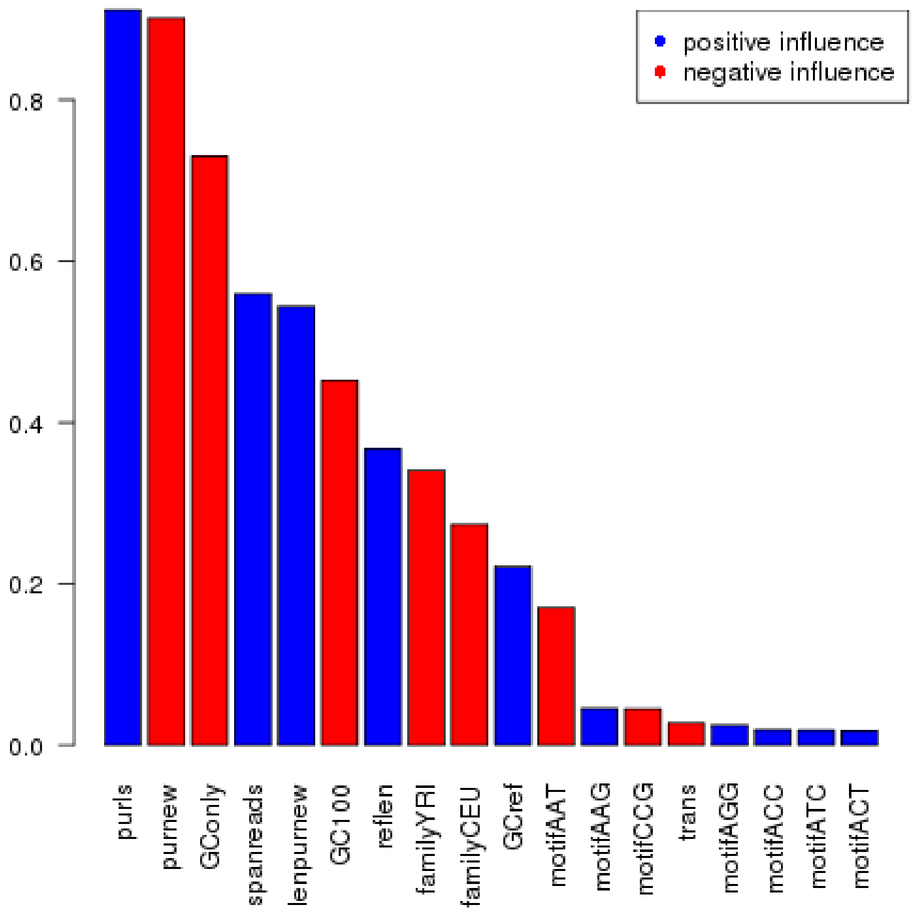

Figure 3.3: Bar graph of absolute values of coefficients from a logistic linear model for a STR being non-reference.

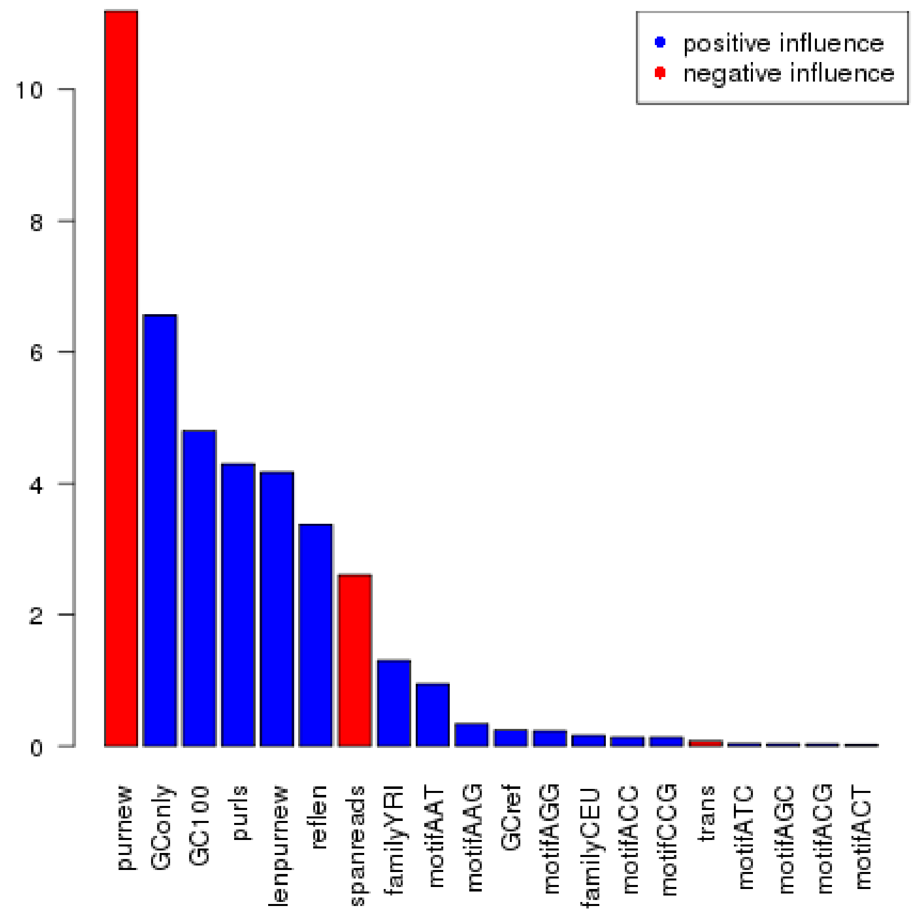

Figure 3.4: Bar graph of absolute values of coefficients from a linear model for the magnitude of an indel at variant STR loci.

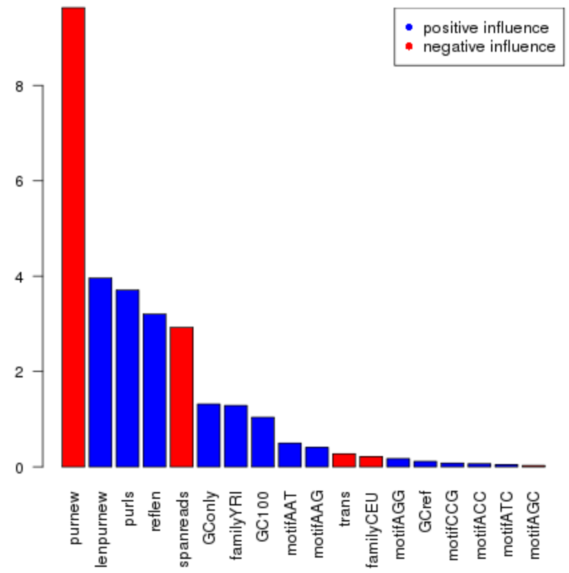

Figure 3.5: Bar graph of absolute values of coefficients from linear model for the magnitude of an insertion at variant STR loci.

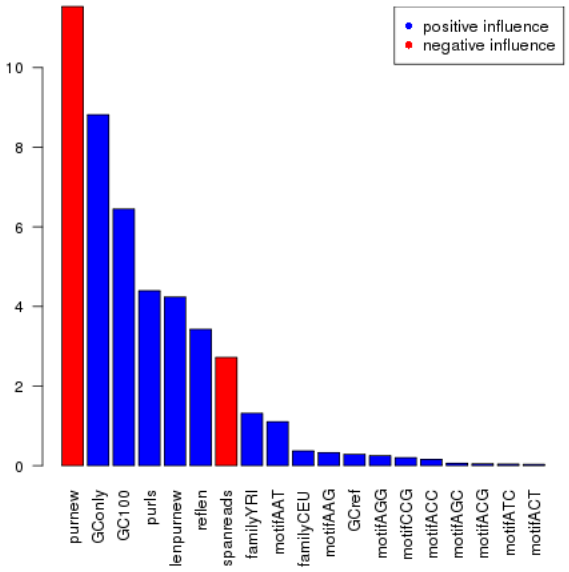

Figure 3.6: Bar graph of absolute values of coefficients from linear model for the magnitude of a deletion at variant STR loci.

3.5.1.1 Bias in modeling of purity

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

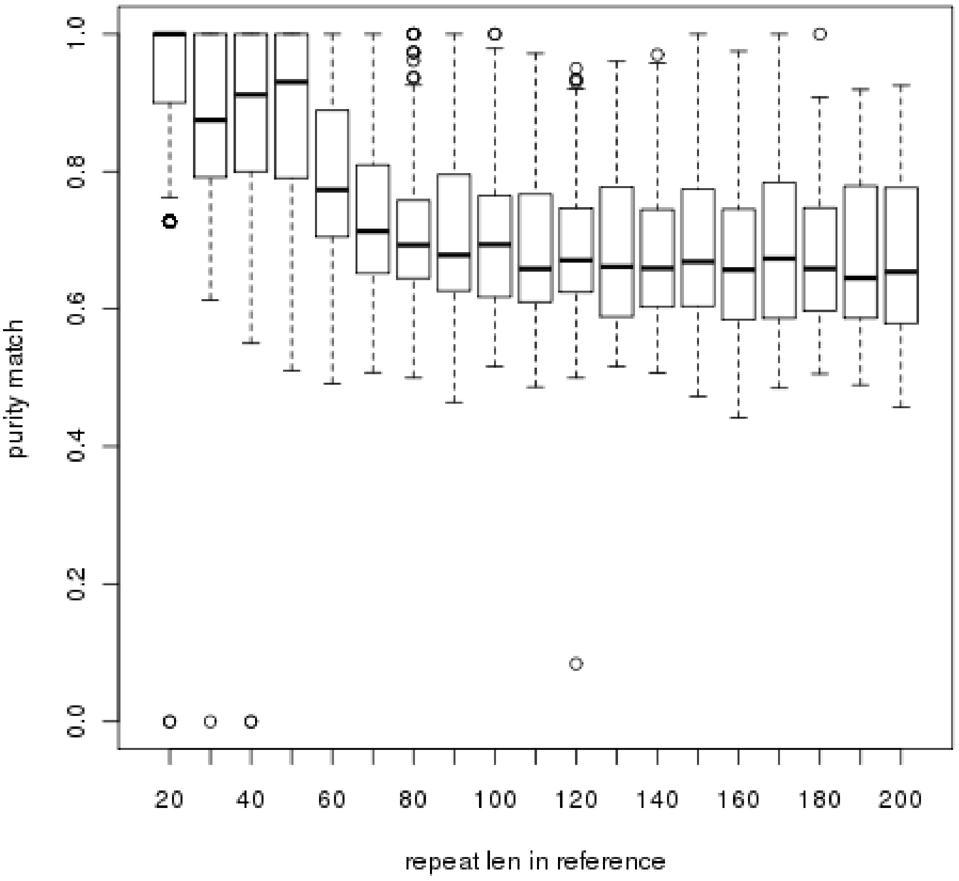

Figure 3.7: Boxplot of repeat purity across varying repeat lengths. This boxplot shows the values for our purity metric described in 3.4.4.2.1.

3.6 Discussion

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

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

3.6.1 Sample family correlations

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

3.6.2 GC composition correlations

Ignoring factors believed to be unimportant biologically or biased (family and length independent purity metrics), what was left were the true set of factors that play some sort of biological role in observing an indel at a given STR locus. Looking at figure 3.3, one of the largest influences on observing a variant is the amount of GC content proximal to the STR locus; the higher this GC content, the less likely you are to observe a variant. It is perhaps surprising that it is the GC composition of the flanking regions rather than the repeat sequence itself that has one of the largest effects overall, as well as the largest amongst the GC content metrics. One technical explanation as to why fewer variants are observed in regions with high proximal GC content is the higher portion of sequencing errors in this region could lower the number of spanning read pairs, in turn lowing the power of our model to detect variants. In order to explore the external factor of mapping bias in the the genome, we compared directly the proximal GC content (GConly) to the GC content within the STR (GCref) based on the number of spanning reads observed at a given locus (see figure 3.8). The difference in the two metrics across the number of spanning read pairs showed that there is an indication that a higher GC content in the proximal sequence is associated with fewer spanning read pairs. For all but two bin sizes (190 and 200, which are the two smallest bins), the amount of flanking GC composition is anywhere from 10% to 35% higher than the STR composition, with the lower spanning read counts showing the strongest bias – which are also the largest bins.

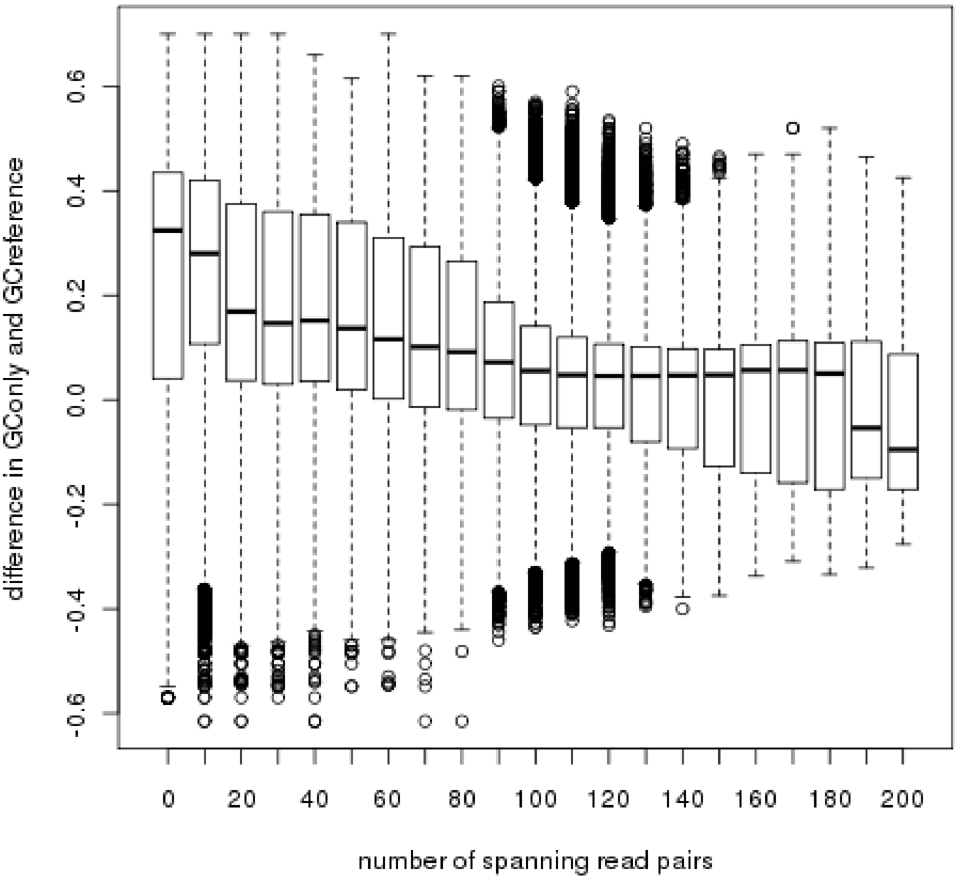

Figure 3.8: Boxplot of differences in GConly and GCref at a locus binned by the number of observed spanning read pairs at a locus. Each bin represents all sites in the genome which have a given number of spanning read pairs independent of the length.

3.6.3 Motif correlations

All but two motifs, CCG and AAT, were positively correlated with observing a variant (compared to the AAC family). While the families AAG, CCG, AGG, ACC, ATC and ACT all have comparable influence compared to one another, AAT has approximately five times more influence than the next strongest family. As the family AAT is the only family to not contain any GC content, this correlation is in agreement with the factor GCref which is strongly correlated in the opposite direction. Most astonishing is while GC composition in the reference is positively correlated with observing an indel, the motif family CCG is negatively correlated. It might be possible that CCG repeats form some sort of secondary structure such as G-quadruplexes which are relatively prevalent in the genome and may decrease the chance of those sites undergoing mutation (Hazel et al. [2004], Huppert and Balasubramanian [2005], Bugaut and Balasubramanian [2008]). This is something to be explored further.

3.6.4 Purity correlations

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

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

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

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

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

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

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

3.6.6.1 All variants

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

The entirety of motif families are also positively correlated. AAT has the largest effect for observing a larger indel but from the previous analysis, is negatively correlated with observing a variant (the largest coefficient of the motif families). This is quite interesting. Motif CCG also exhibits this interesting reversal.

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

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

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

3.6.6.2 Independent analysis of insertions and deletions compared to the reference

When comparing the magnitude of indel calls in insertions versus deletions, almost all correlational directions match one another with the exception of the motif family AGC which is negatively correlated in inserts and positively correlated in deletions. Its effect, however, is relatively small in both directions and is most likely statistically insignificant. The correlations that stand out the most are in the same relative order of significance. While the strongest indicators of larger variants are the purity metrics for insertions, it is the proximal GC content for deletions. All other factors seem to be in the same order and relative influence to one another. A simple explanation is not readily available and warrants further analysis. It should also be noted that the reference genome does not represent the ancestral state. Many of the tandem repeats were estimated using BACs and so at variable loci the allele present in the BAC was chosen, which typically will represent a selection at random according to the population allele frequencies. This makes inference difficult when comparing whether insertions or deletions are more likely as we can not say for sure that the alleles in the reference represent the ancestral state.

3.7 Conclusion

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