Chapter 4

Population based analysis of short tandem repeats

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

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

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

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

4.1 Low coverage individuals in the 1000 Genomes Project

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

4.1.1 Sources of sequence

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

4.1.2 Sequencing statistics

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

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

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

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

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

4.1.3 Population MPERS distributions

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

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

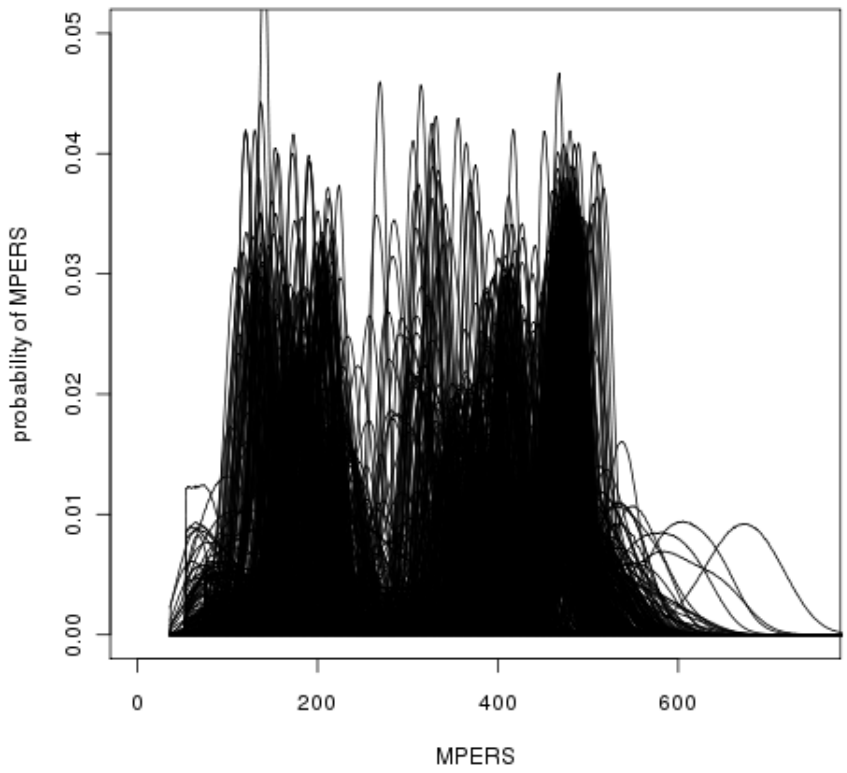


Figure 4.1: Plot of MPERS distributions for every library in the 1000 Genomes Project data set. Most libraries' MPERS fall within the range of 150 to 600 bp with peaks (fragment library sizes) around 150, 200, 400 and 500 bp.

4.2 Modeling

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

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

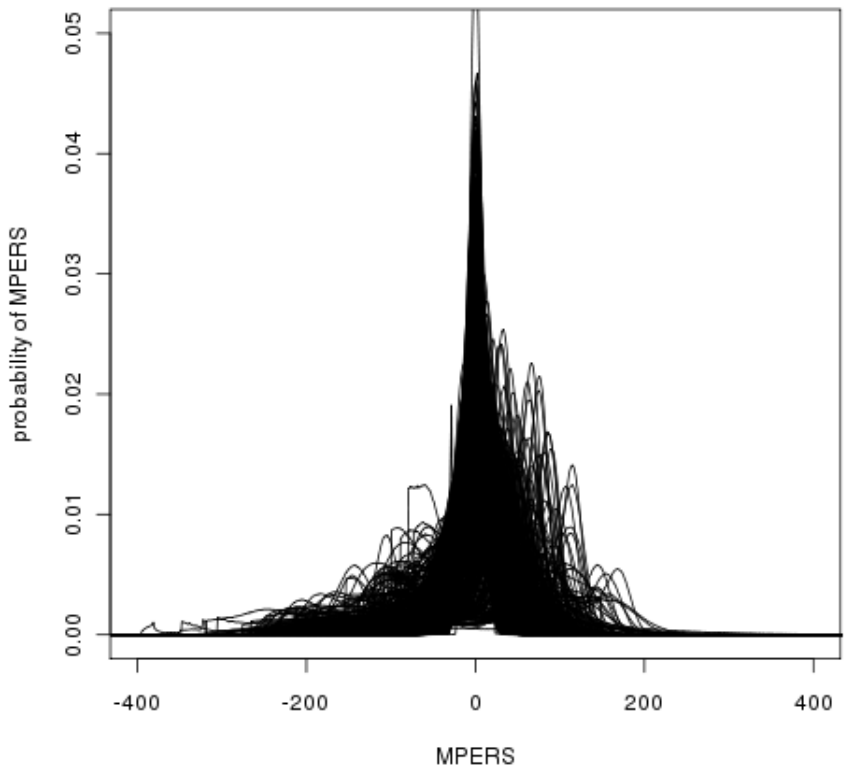
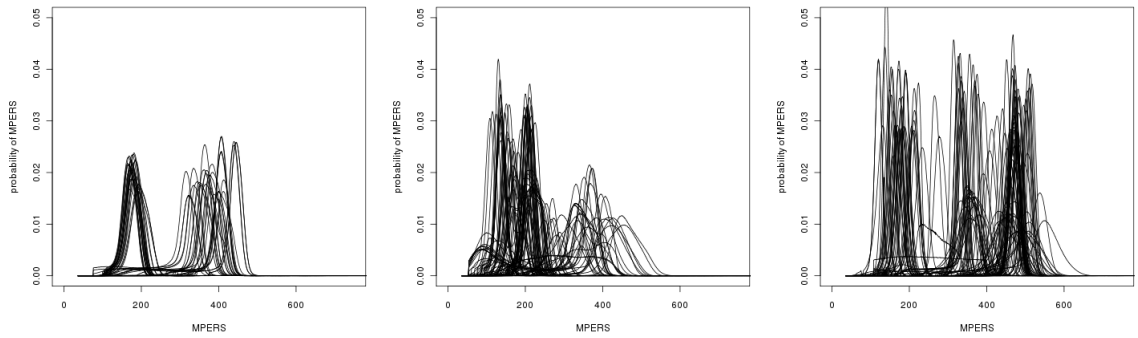


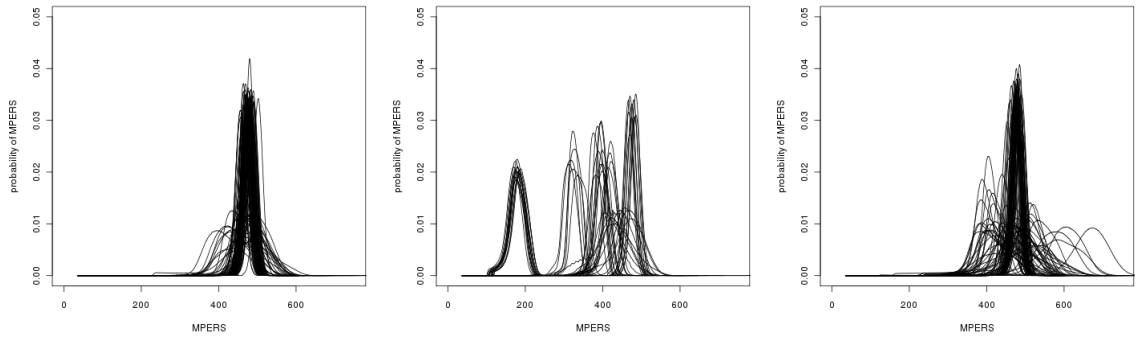
Figure 4.2: Plot of MPERS distributions whose mean of each library is arbitrarily set at zero. It is clear that the majority of libraries have a MPERS variance that is tightly bound around the mean value (peak at zero). This does not mean that all libraries behave well (as signified by the MPERS distributions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

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

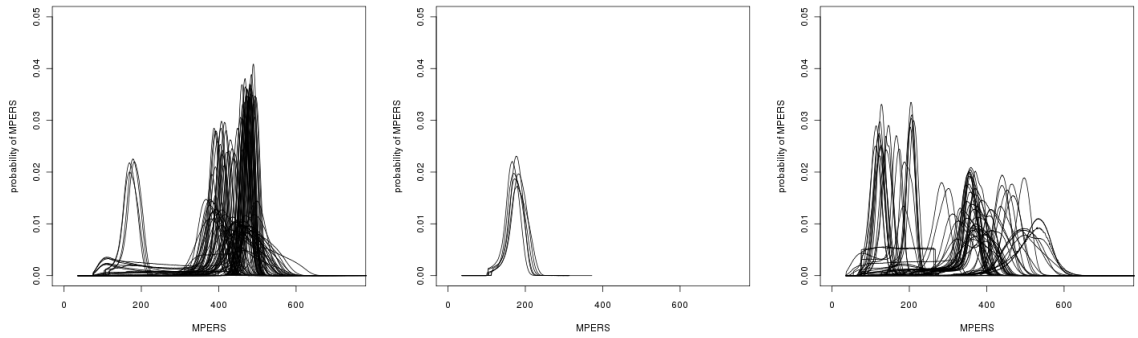
Not interested in a specific individual's genotype, our previous prior over diploid genotypes from chapter 2 was no longer appropriate. Instead, we needed a prior over all distributions of alleles at a locus. As we were no longer looking for genotypes, but allele frequencies, it meant that our posterior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



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

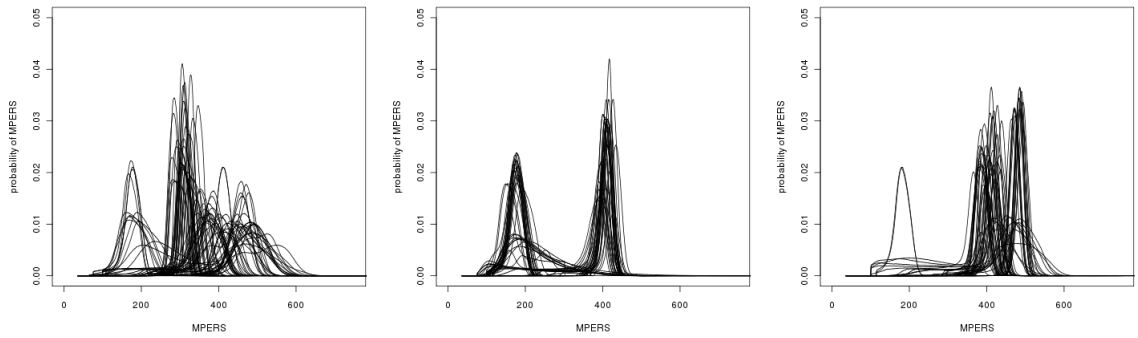


(d) MPERS for CHS population (e) MPERS for CLM population (f) MPERS for FIN population

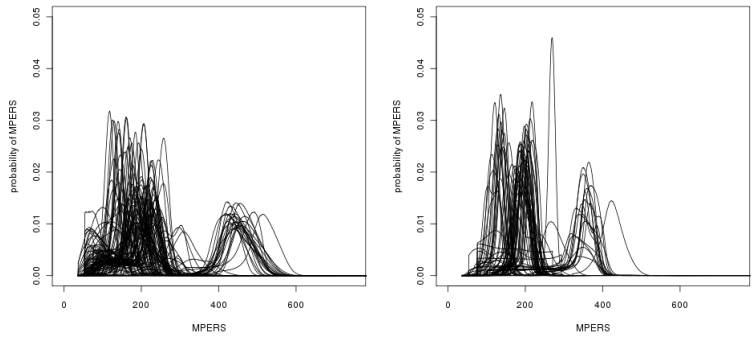


(g) MPERS for GBR population (h) MPERS for IBS population (i) MPERS for JPT population

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



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



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

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

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

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

has a probability density function given by

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

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

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

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

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

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

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

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

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

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

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

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

4.2.1 Priors

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

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

4.2.2 EM algorithm

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

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

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

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

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

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

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

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

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

which in log space would be

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

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

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

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

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

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

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

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

4.2.3 Gibbs sampling

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

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

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

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

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

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

4.3 Simulation

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

4.3.1 Simulation of MPERS for spanning read pairs

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

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

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

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

4.3.2 Simulation results

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

4.3.2.1 Reference allele frequency

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

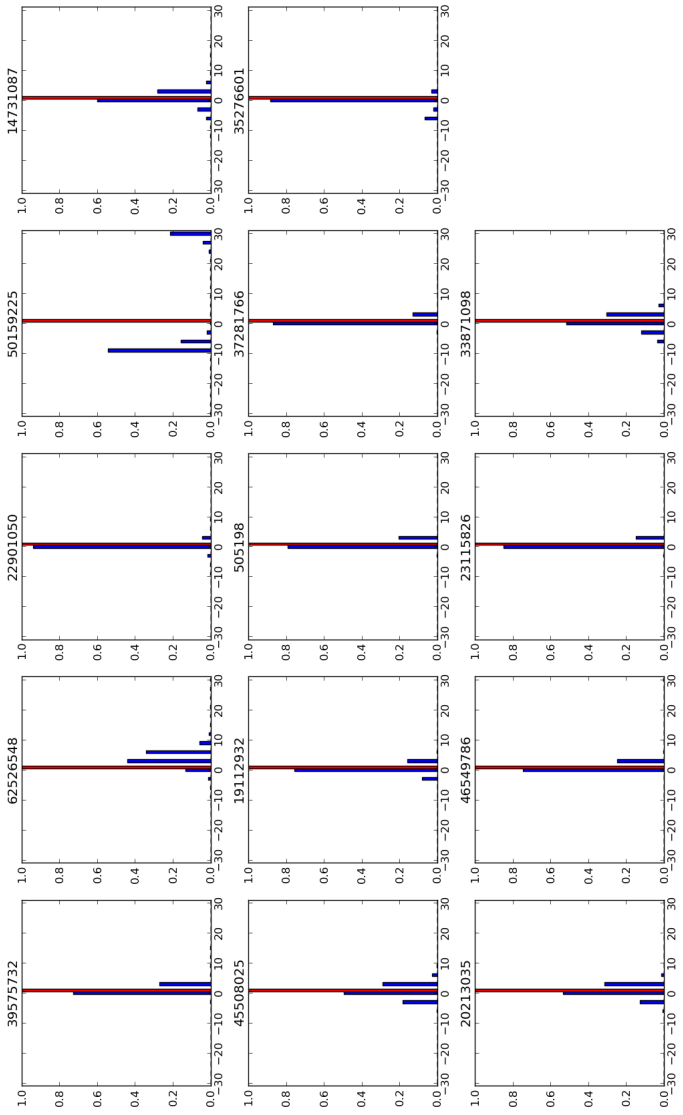


Figure 4.4: Prediction of allele frequency distribution for the EM algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely of reference alleles based on a CHS population (red bars). Most all the predictions' allele frequency distributions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

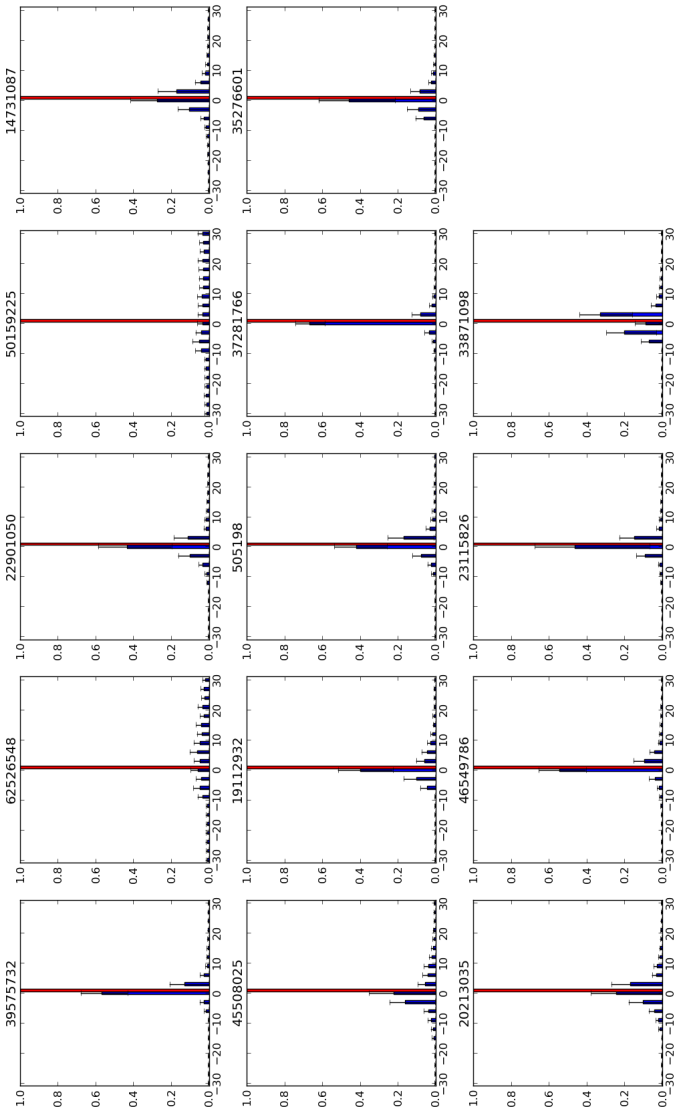


Figure 4.5: Allele frequency distribution prediction of alleles for the Gibbs sampler algorithm (blue bars) in 14 simulated loci in chromosome 20 from an underlying allele frequency distribution comprised solely of reference alleles based on a CHS population (red bars). Most of the predictions' allele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

4.3.2.2 Two and three allele population frequency alleles

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

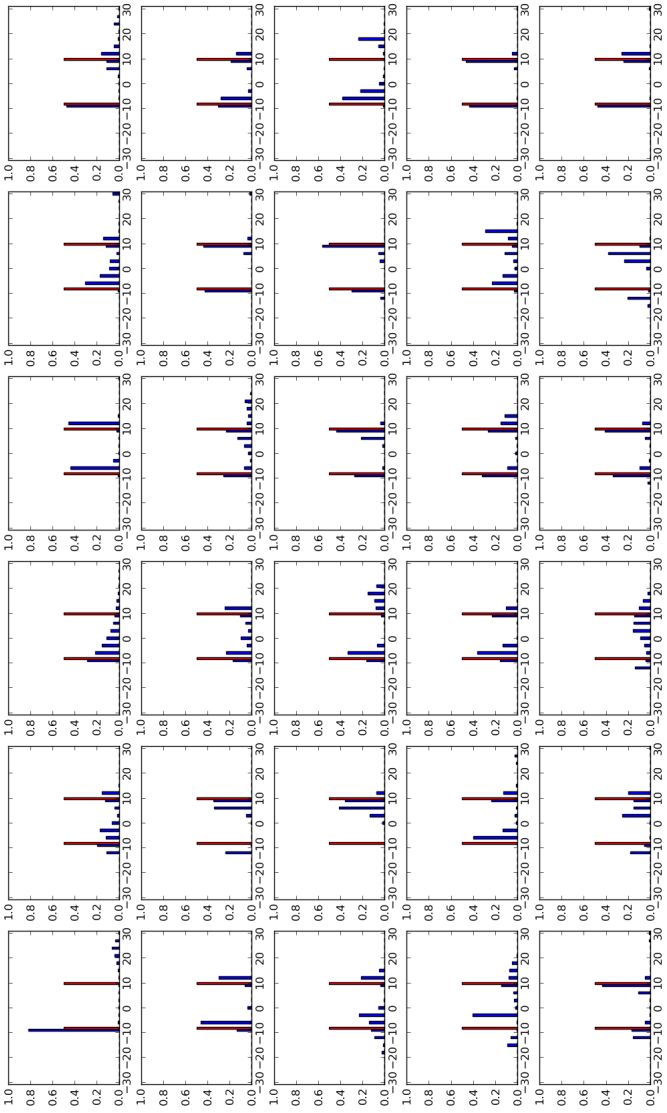


Figure 4.6: Allele frequency distribution prediction of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency distribution of ±9 bp each at a 0.5 frequency (red bars) based on a CLM population. As with the reference simulation, the EM is much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

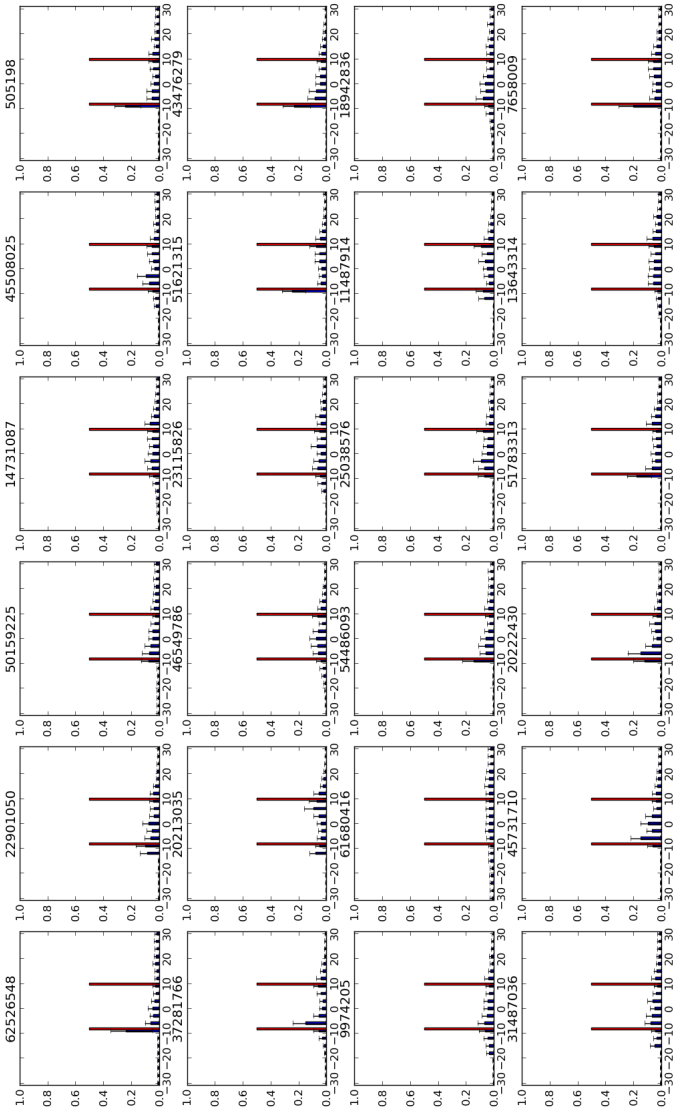


Figure 4.7: Allele frequency distribution prediction of alleles for the Gibbs sample algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of ±9 bp each at a 0.5 frequency (red bars) in a CLM population. Not as aggressive as the EM, sites show lower frequency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

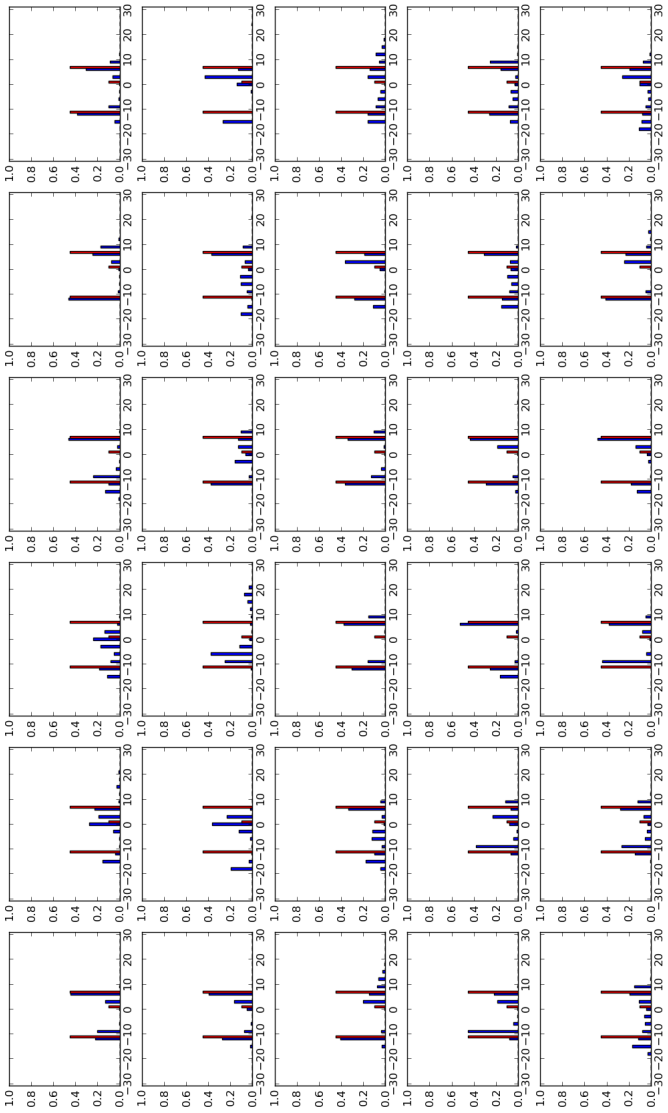


Figure 4.8: Allele frequency distribution predictions of alleles for the EM algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based on a CLM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

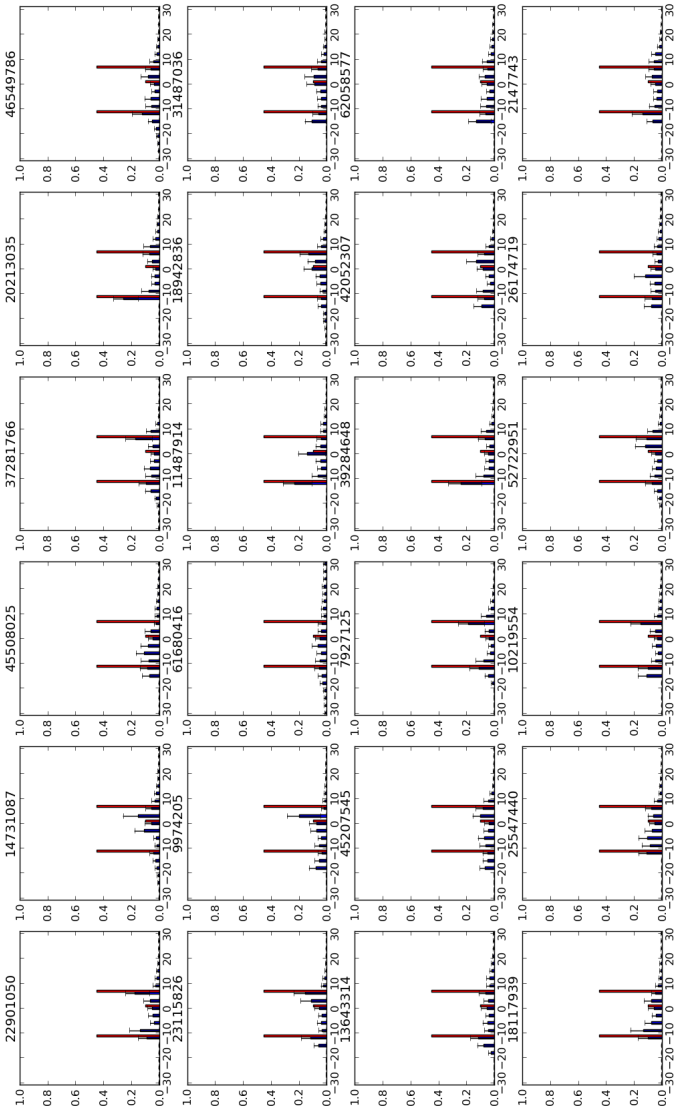


Figure 4.9: Allele frequency distribution predictions of alleles for the Gibbs sampler algorithm (blue bars) in 30 simulated loci in chromosome 20 from an underlying allele frequency of 0.45 at both -12 bp deletion and 9 bp insertion alleles and a 0.1 frequency at the reference allele (red bars) based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

4.3.3 Simulation results comparisons

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

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

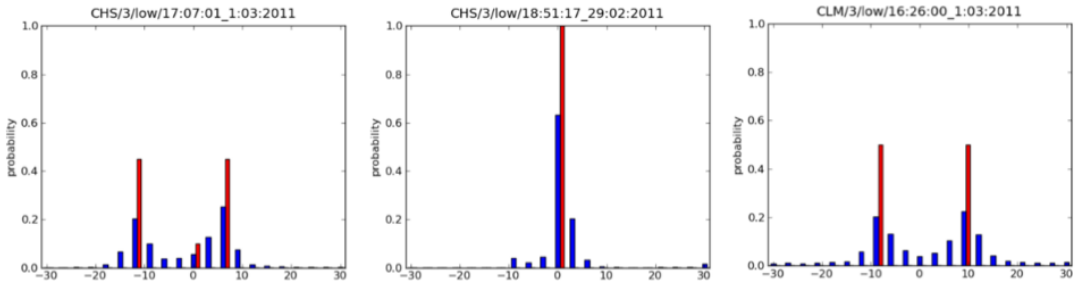


Figure 4.10: Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for the EM algorithm.

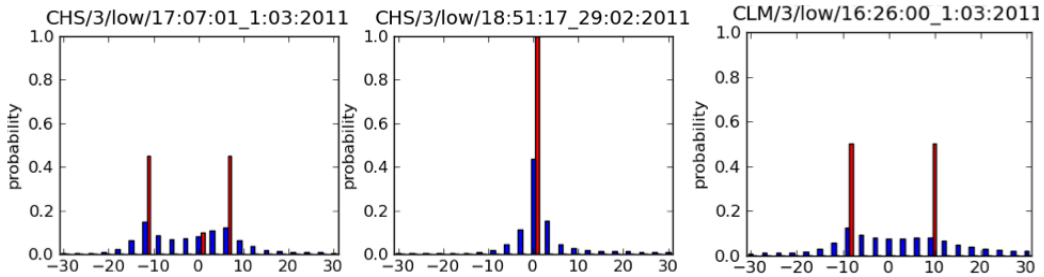


Figure 4.11: Averages of allele frequency distributions (blue bars) across chromosome 20 for three simulation scenarios (red bars) for Gibbs sampling algorithm.

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

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

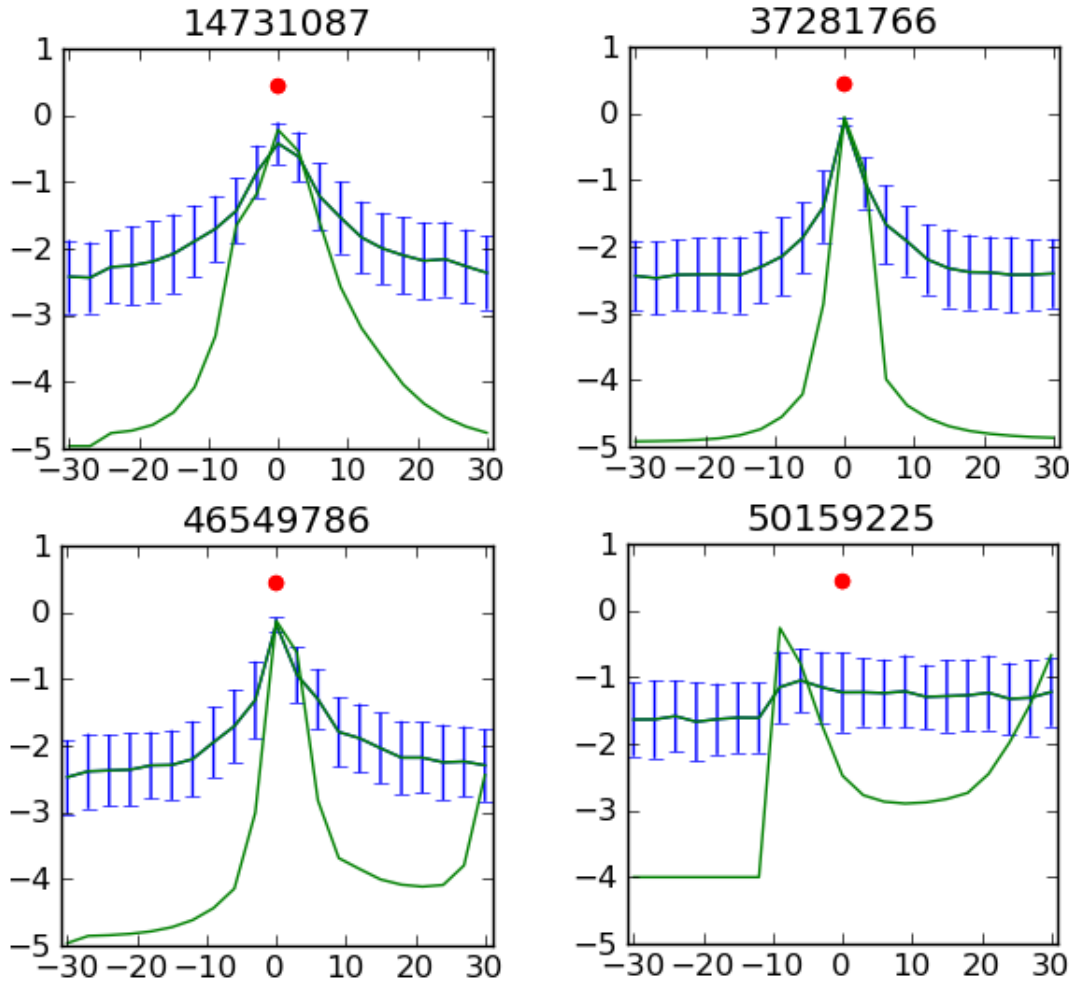


Figure 4.12: Comparison of the EM and Gibbs sampler algorithms for a reference allele frequency distribution. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying allele. The sole green line represents the values for the EM, while the green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

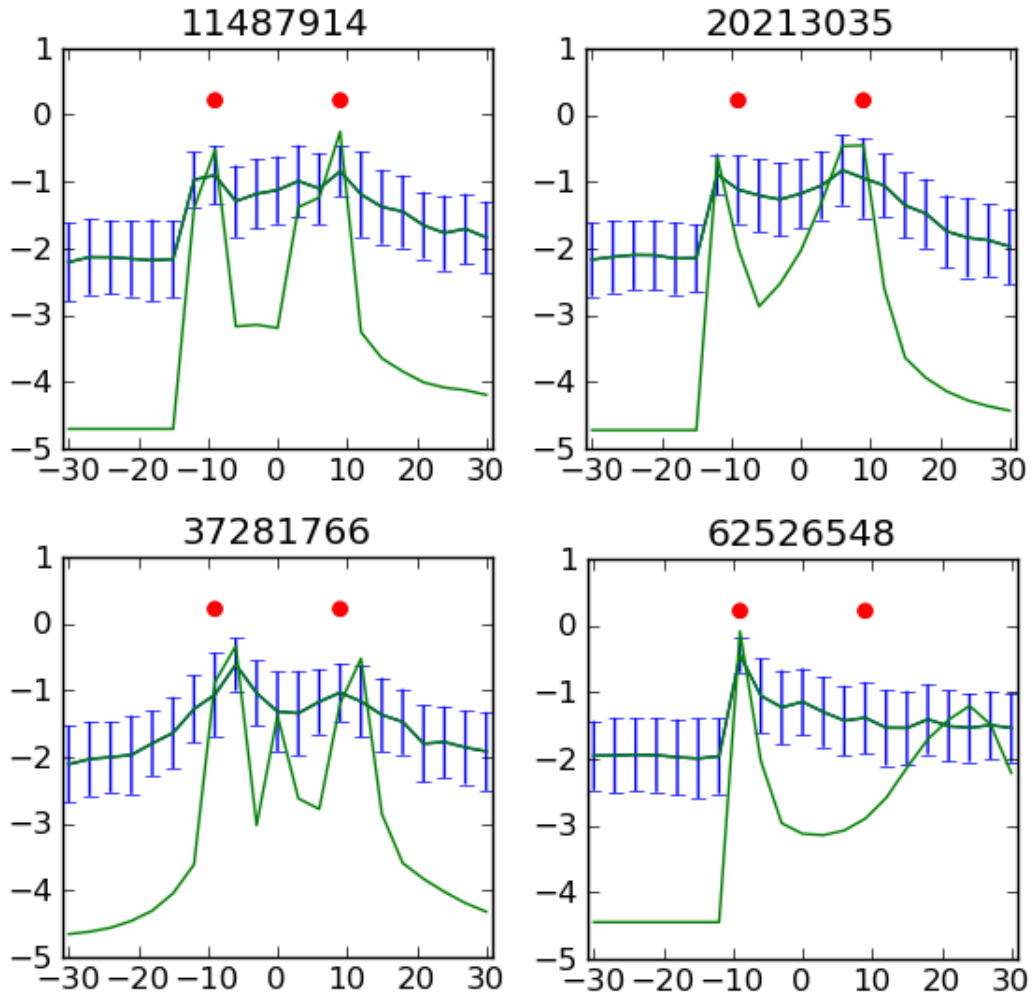


Figure 4.13: Comparison of the EM and Gibbs sampler algorithms for a two allele frequency simulation. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying alleles. The sole green line represents the values for the EM while the green line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

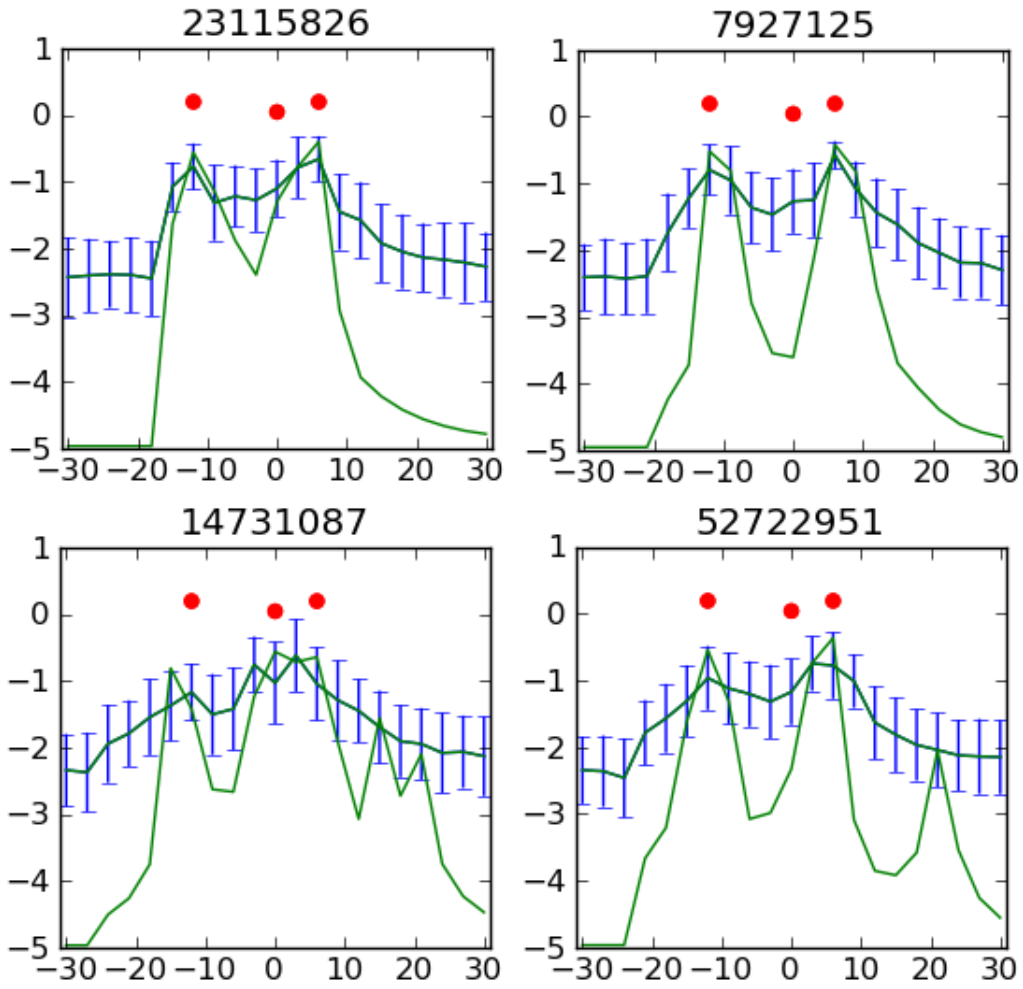


Figure 4.14: Comparison of the EM and Gibbs sampler algorithms for a three allele frequency simulation. The y-axis is the log probability of the frequency of a given allele with the red dots denoting the true underlying alleles. The sole green line represents the values for the EM while the green line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

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

4.3.4 Test statistics

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

4.3.4.1 Entropy

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

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

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

4.3.4.2 Off reference/± 3 bp

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

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

4.3.5 False discovery rate

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

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

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

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

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

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

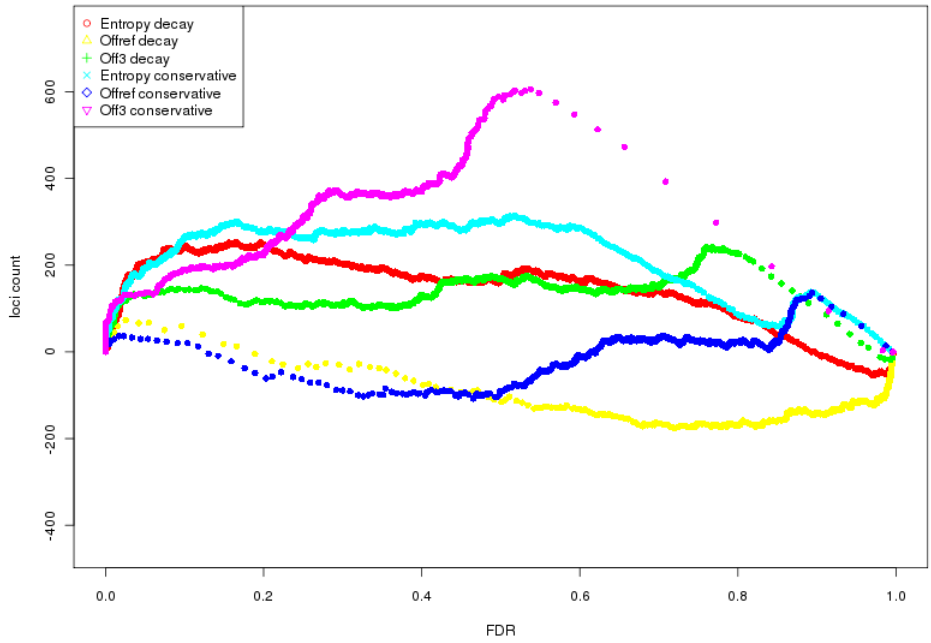


Figure 4.15: Plot of FDR versus true calls for the ASW population for triplet repeat loci on chromosome 20.

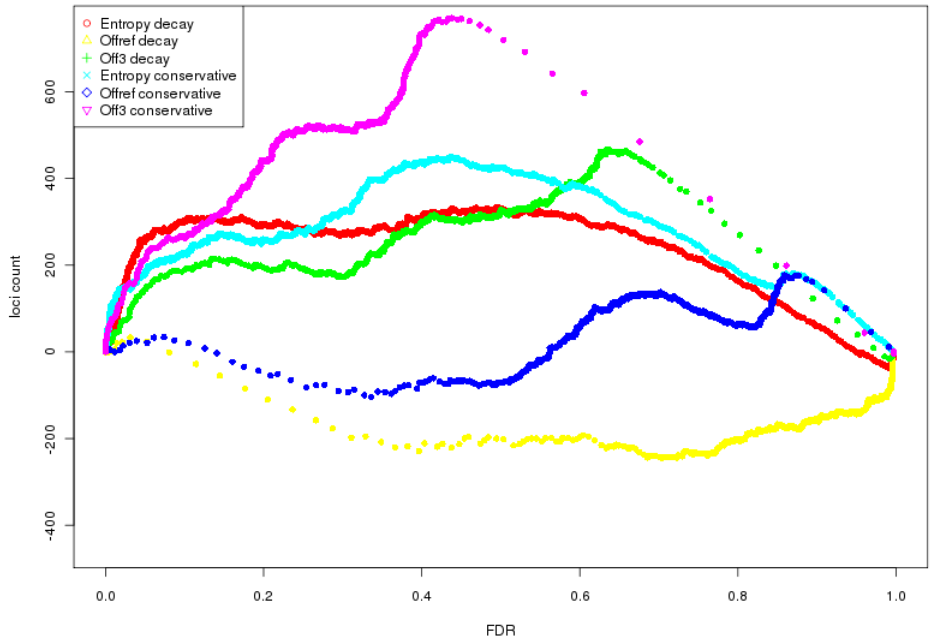


Figure 4.16: Plot of FDR versus true calls for the MXL population for triplet repeat loci on chromosome 20.

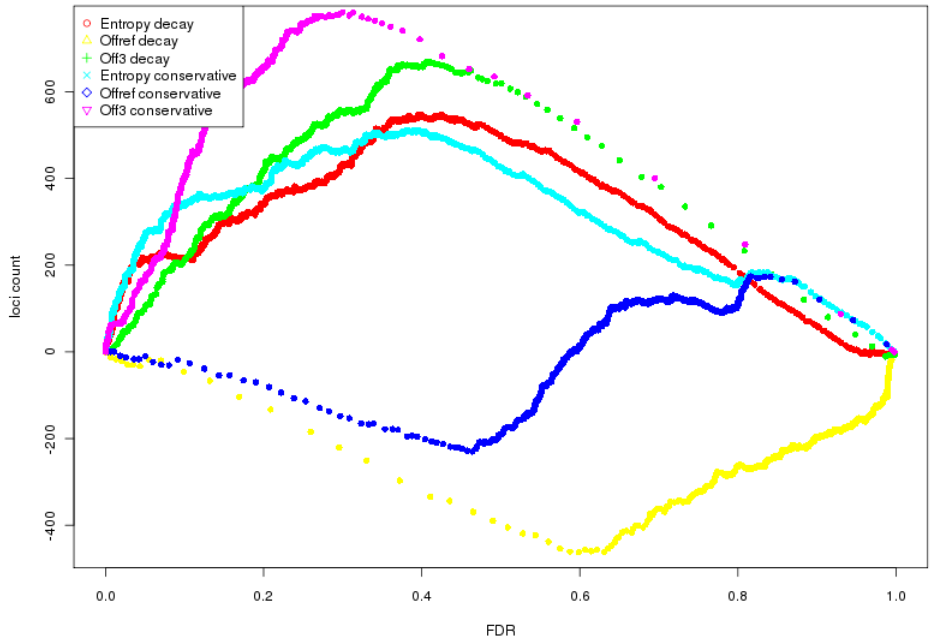


Figure 4.17: Plot of FDR versus true calls for the PUR population for triplet repeat loci on chromosome 20.

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

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

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

4.4 Results

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

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

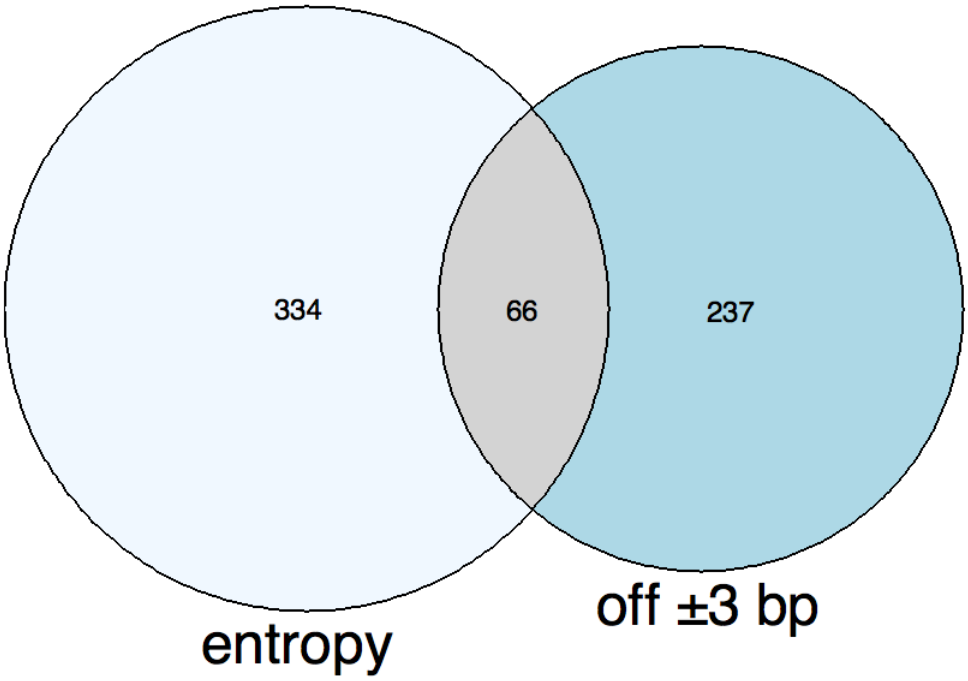


Figure 4.18: Venn diagram of intersection of significant loci called by entropy and off ±3 bp.

4.5 Discussion

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

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

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

4.5.1 Factors

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

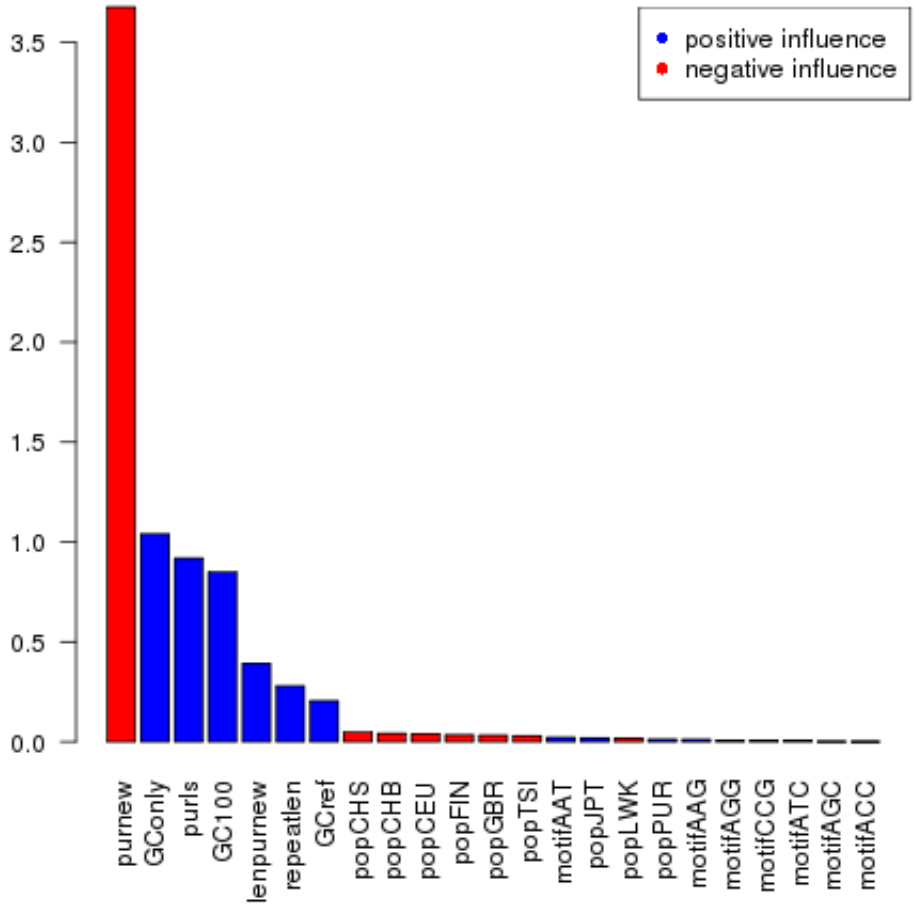


Figure 4.19: Bar graph of absolute values of coefficients from logistic linear model on whether a locus's entropy value is significant against various factors.

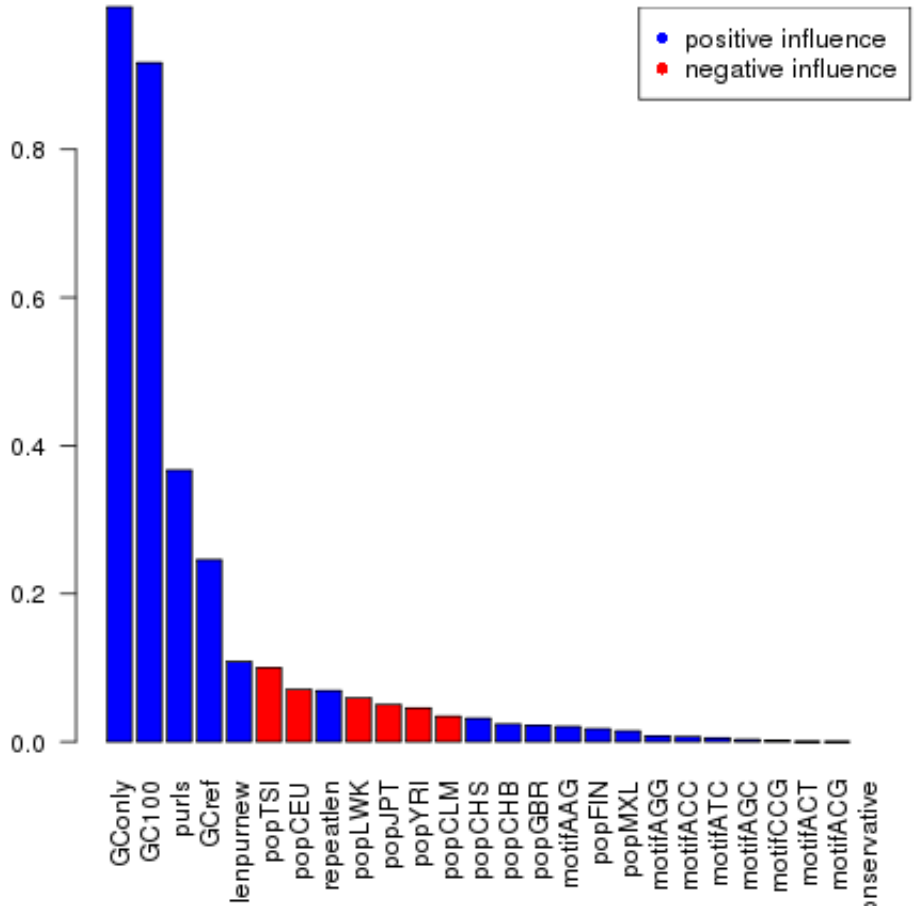


Figure 4.20: Bar graph of absolute values of coefficients from logistic linear model on whether a locus's off ±3 bp value is significant against various factors.

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

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

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

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

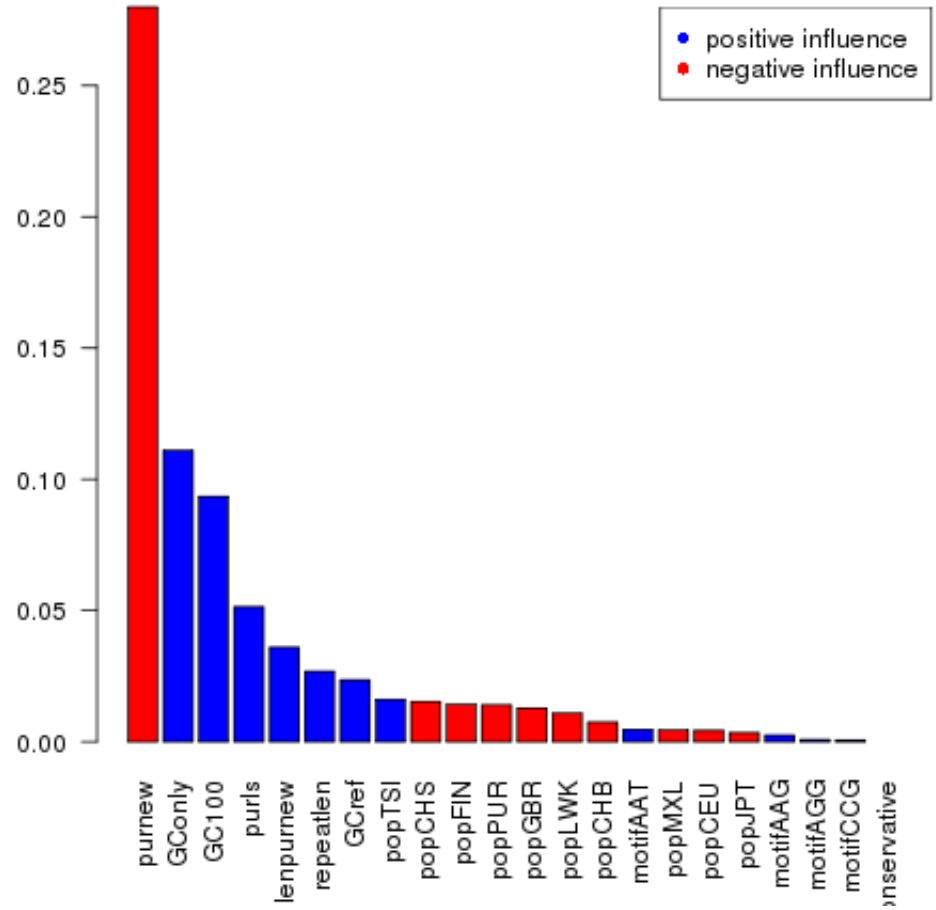


Figure 4.21: Bar graph of absolute values of coefficients from linear model of significant entropy loci values and the various explanatory factors.

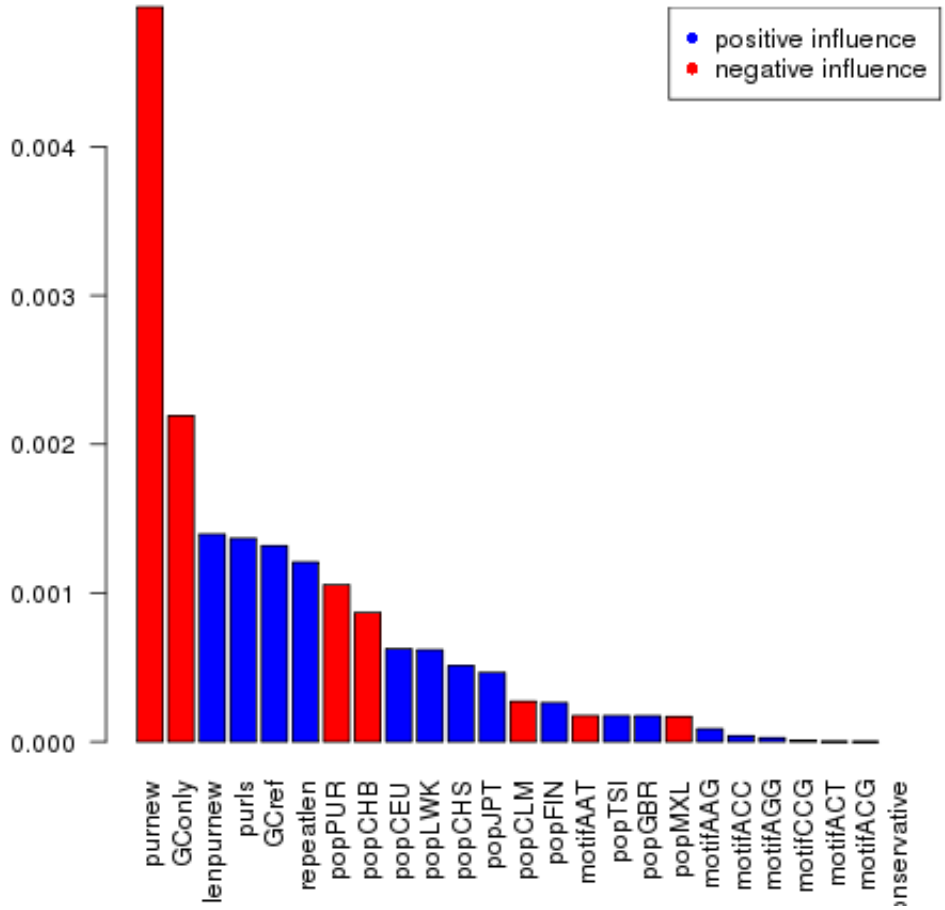


Figure 4.22: Bar graph of absolute values of coefficients from linear model of significant off ±3 bp loci values and the various explanatory factors.

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

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

4.6 Conclusion

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

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

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