Chapter 5

Conclusions

5.1 Conclusions, discussion and future work

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

5.1.1 Modeling variation in STRs

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

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

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

5.1.1.1 Future work

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

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

5.1.2 Characterizing STR variation

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

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

5.1.2.1 Future work

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

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

5.1.3 Modeling STR loci in large population data sets

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

5.1.3.1 Future work

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

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

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

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