Chapter 5

Conclusions and Future Work

5.1 Conclusions

The application of association genetic methods to cellular traits is an important area of research with great potential to reveal new information about cellular functions, and help interpret the genetics of diseases and other whole organism traits. This thesis contributes to the understanding of how DNA variation regulates transcription factor binding by investigating a key transcription factor CTCF. In chapter 2, I described a study that performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for lymphoblastoid cell lines collected from 51 HapMap individuals. We have identified tens of thousands of bindings sites with the vast majority showing large inter individual variations. Most of the binding sites are identified with a matched CTCF canonical motif, reflecting sequence specificity of the transcription factor. To reveal the genetic contribution to such variation, we performed QTL analysis using a linear additive model, and focused on *cis* regulation within 50kb genomic window. Our results suggest a strong genetic basis for variation at many binding regions. The distributions of the physical locations as well as the corresponding effect sizes of the QTL variants show clear tracking on the information content with the motif, which suggests that mutations at the DNA-protein binding interface exert a functional impact on binding proportional to the predicted binding affinity of the nucleotide. It

also shows strong correlation with sequence conservation, which supports the functional impact of the variants. Interestingly, there are binding regions whose variation of binding intensity appears to be only attributable to variants that are distal to the sequence motif. We recognized that in a substantial fraction of these cases there are variants within the motif but with low frequency (<5% MAF). Particularly, these low frequency variants are often in linkage disequilibrium with the lead QTL variants that are distal, indicating a phenomenon that the lead QTL variant is actually tagging a putative functional on-motif variant but with lower frequency. Taken this together, we show that the majority of CTCF binding QTLs have genetic regulatory variants in close vicinity.

Also in this work, we suggested a novel discovery of CTCF binding patterns on the X chromosome (See Ding et al., 2014 for details). Particularly we showed there exist three different types of binding in regions on the X chromosome: ones that only bind on the inactive X; ones that bind on both X; and ones that only bind to X in females. Also CTCF binding shows much stronger correlations between nearby sites on X than on autosome, suggesting important roles of CTCF in X inactivation. Together these suggest that CTCF maybe involved in a large scale chromatin remodeling associated with inactivation, although the specific functions of the different types of sites are yet to be found.

The resulting data from this study also contribute to the growing cellular information accumulating around the HapMap cell lines. Currently most public data on transcription factor binding are on a handful of samples, quite often only one cell line. This study provides a precious resource for scientists to try to understand the global picture of cellular regulation by looking at both intra-individual and inter-individual variations. Statistical methods that adjust for noise in cellular assays may also be able to use it to better model noises in transcription factor binding assays.

In chapter 3 we tackle the problems of performing phenotype association using noisy gene expression data. The expression levels of functionally related genes, such as genes in a signaling pathway, tend to correlate to each other for a biological reason. Taken this common variation, feature extraction methods can effectively extract signal out of noise present in data from individual genes. We applied factor analysis firstly

to remove systematic noise in the entire gene expression data, then to each individual pathway to extract pathway factors as our new phenotypes. We showed that these pathway factors are substantially more heritable than individual genes, using estimates from the twins data in our study. Using ageing phenotypes, our approach revealed a number of pathways known to be related to ageing as well as new pathways that are candidates for further investigation.

In this thesis, I also worked on developing new methods for important cellular traits. In chapter 4, I described a novel method that uses whole genome as well as exome data to estimate telomere length. It is particularly attractive to the large sequencing cohorts generated from cancer, epidemiology and ageing studies. Many important questions about the role of telomere length regulation can be tackled as the phenotype data can be made available by the new method. I have already applied it in a melanoma study which discovered that mutations in POT1 gene predispose to the disease and the mutated POT1 gene is associated with longer telomeres.

5.2 Future Work

Genomic DNA stores functional information that makes diverse biological systems in various cells and organisms. In addition to the protein-coding sequences, there is also a group of sequences that can indirectly influence the expression of genes as regulators. These sequences are key in achieving the complexity of the organisms. For example, having hugely different number of cells and cell types, *Caenorhabditis elegans* (around a thousand cells) and humans (trillions of cells) have quite similar amount of protein-coding genes (around 20 to 25 thousands).

A large volume of research efforts have been made to understand the regulatory mechanisms. The genomic regions that are immediately upstream of the transcription start sites have been found critical for controlling gene expression (See review Wittkopp and Kalay, 2011). In addition to these regions, one important phenomenon of gene regulation in higher order organisms is that the regulatory machinery needs not always be proximal. Increasing amount of studies have shown that there exists functional elements that are spatially distal to genes but are capable of influencing gene expression

(See review Bulger and Groudine, 2011). This is possible as chromatin is in a three dimensional space, where elements that are distal to their target genes can be brought close by higher order structures. CTCF is one of those special proteins that is involved in such function. Systematic studies on its bindings including this thesis show that its binding is regulated by genetic factors. Although nearly half of the QTL variants are close to the canonical motif, many are not. Only a minority of QTL variants are within the binding motif affecting binding directly. It is possible that the genetic effects of the proximal variants are mediated by collaborative factors. These requires investigation of binding patterns of more transcription factors in a group of individuals to search for correlation of bindings between them. Also, importantly, a higher resolution map, ideally at single base pair, is needed for identifying such interactions by knowing exactly where they bind. The current standard ChIP-seq experiments produce binding peaks at a resolution of a few hundred base pairs, much larger than the binding interface. Methods such as the enrichment analysis used in this case may not be reliable for individual events. Some of the recent technologies, such as Capture-C (Hughes et al., 2014), shows some promising directions by producing base pair resolution for interactions between a pair of factors in a *cis* window. Meanwhile, it is also possible that transcription factors not always bind to their canonical motifs, but is subject to a probability as a stochastic process. The bindings may occur at sites with weaker motif pattern that are yet to be discovered. This could be more confidently identified if we know the exact positions of binding, which gives much better signal noise ratio than searching for motifs in long sequences.

Future studies should also extend beyond lymphoblastoid cell lines to more tissues such as neurons, muscles etc and attempt to collect measurement *in vivo*. By doing so the chance of revealing the genuine biological mechanisms are much elevated. The diversity of regulation in these tissues provide a natural platform for comparing binding patterns of transcription factors and expression patterns of genes. Associating these patterns with the developmental features of these tissues may give important information on the regulatory mechanisms.

In this and many other studies, QTL analysis and allele specific analysis are done separately. This is because *cis* regulation patterns can be captured both by comparing

between individuals, where individuals with AA, AB and BB genotypes should have different phenotypic levels if the locus is causal, or comparing within individuals, where in individuals with AB genotype the signals from A and B alleles should be different. This can be combined into a joint haplotype test that increases the statistical power (see methods in McVicker et al., 2013). Additionally, quite often the loci showing an allele specific signal are not the causal loci themselves, supported by the observations of conflicting signal directions of allele specific alleles between different individuals. Some initial methods have been developed to search in the local region to find variant that maximize a test score for consistent allelic imbalance (Lappalainen et al., 2013). Methods that extend towards these two directions only just begin to emerge and there is much rooms to develop them further. One approach could be parametrize phenotype and genotype value by haplotype instead of by individual. An individual that is homozygous and has a phenotypic value of 10 would be encoded as two entries in a (genotype, phenotype) format as (0,5), (0,5) or (1,5), (1,5) depending on whether the genotype is 0 or 1. And an individual that is heterozygous would be encoded as (0, allelic value for genotype 0) and (1, allelic value for genotype 1), taking the actual measured allelic value from these individuals. Associations can be tested using a linear model linking the two variables. This way the inter-individual QTL test and the intra-individual allele test can be combined. One caveat of such encoding is that the two entries of an individual are not independent. This is however similar to correcting for population structures, such as for monozygous twins. One could use a linear mixed model to correct for such structure using a relationship matrix.

As more and more functional data are accumulated around study samples, ideally one would want to model them together, looking for not only the effect of genotypes on each individual phenotype, but also on the network of them, learning their interactions and eventually the causality directions. Taking advantages of improved phasing algorithms, studies have started to investigate the phenotypes at a haplotype level, which extends to a larger genomic regions beyond a single SNP, and allows to model many other phenotypes that occur on the same haplotype. For example, the CTCF data generated in chapter 2 can be combined with RNA-seq data published for the same individuals for such analysis. One could model the allelic effects of CTCF binding and

gene expression using two binomial variables, and see if they behave independently. Interestingly, haplotype that links two traits can be broken by recombination, forming recombinant samples. This gives an opportunity to distinguish the causal signals if any, because if the direction of association appear both in non-recombinants and recombinants, it suggests that two traits are causally linked.

The algorithm and software developed in chapter 4 have already generated strong interest in cancer and ageing studies. I am involved in two ongoing cancer studies on prostate cancer and melanoma while I am writing this thesis. In the ageing context, some recent work shows strong parental age effect, particularly paternal effect, on the health and disease status of their offspring (Kong et al., 2012, Goriely and Wilkie, 2012). Such paternal effects can be associated with telomere lengths as longer telomeres are transmitted to offspring by older fathers. Telomere lengths can be estimated from trio sequence data that some already become available from epidemiology studies. The age at conception can be easily worked out if the ages of the trio are known. It is also of interest to understand the heritability of telomere length, which can also be worked out using these data. Additionally, genome wide association analysis can reveal the genetic loci that are associated with the variation of telomere lengths, which will be very relevant to the context of ageing, cancer and a number of other diseases.