Chapter 1

Introduction

Life presents enormous diversity. From the white snow flower growing in the over 2000 meters high plateau in Tibet to the lionfish swimming in the Indo-Pacific sea, each life form has dramatically different appearance, structure, behavior, reproduction etc. Yet they fit in a global ecosystem finding their own ways of living, with their positions and roles shaped by the force of evolution. Fascinatingly, given how distinctive each life is, there are things that are shared and principles that they follow. Understanding how diverse life arises and how the characters transmit and spread between generations and species is key to reveal the basic principles that govern their biology.

Some of the patterns must have been realized by ancient humans. These include that children are more likely to have similar appearance to their parents in having similar eye colors, skin colors, height and so on. However, it was not at all straightforward to understand the reason for such similarity between parents and their offspring. To answer this question, a number of prerequisite questions need to be addressed first. How is information faithfully maintained in individuals, and the cells within them? How is this information transferred from one generation to the next and how does it control the characteristics of an individual?

Here I briefly review the progress in the genetics of traits that are at the level of individuals, and how we can use a similar approach to study the genetics of molecular traits at a cellular level, which must come between the genetic material and the

organismal trait. I firstly introduce the history of mapping Mendelian traits and quantitative traits. I then discuss widely used methods that have been developed to date for genetic mapping. Finally, I describe the progress in studying cellular traits using the same principles.

1.1 Hunting for genetic determinants of phenotypes

1.1.1 Mendelian traits

The understanding of the basic principles of genetics has come a long way in the last 150 years. In early and mid nineteenth century, the use of hybridization in plants to obtain flowers with desired colors was already studied scientifically (Gärtner, 1849). But the principles that govern the formation of the traits were yet to be formulated. Gregor Mendel chose peas (*Leguminosae*), which has clearly distinguishable characteristics and good protection at flowering time from foreign pollen contamination, for his studies(Mendel, 1865). After eight years of counting the number of peas with different seed coat colors, shape etc, his milestone paper in 1865 illustrated the principles that were later referred to as Mendelian Laws, which became the corner stones of the genetic field.

During the same period, scientific progress on cytology discovered possible physical molecules or structures that can be linked to Mendelian factors. In 1866 Ernst Haeckel postulated that the nucleus is responsible for heredity from the observation that sperms largely contain nuclei. Deoxyribonucleic acid (DNA) was first isolated by a Swiss physician Friedrich Miescher in 1869 and later in 1875 Strasburger discovered chromosomes. Sutton and Boveri in 1903 proposed the “chromosomal theory of inheritance”, which suggested a direct link between the Mendelian factors and a physical cellular molecule. The inheritance material was confirmed later in 1952 in the famous bacteriophage experiment by Alfred Hershey and Martha Chase.

1.1.2 Quantitative traits

Mendelian factors can be observed in traits that are separated into clear categories, such as the color of seed coats. But there are also, perhaps more prevalently, traits that are continuous and not clearly separable into discrete classes, such as the weight of peas or the height of plants. Many of these traits are also highly heritable. Initially, there seemed to be little connection between Mendelian factors and continuous traits. Early scientists were unable to discover a simple rule of heredity in these traits. Breeding studies such as East 1916 suggest that absolute dominance is rare. Even a Mendelian trait such as the plant color, with careful examination, still shows some variation. This suggests that quantitative characteristics probably result from the action of the environment on the segregation of many Mendelian loci. Statistical developments at the same time were helpful in reconciling the disconnection. Fisher's paper (Fisher, 1919) first introduced variance decomposition, which mathematically illustrated that the variance of a trait can be separated into different components, including those driven by genetic factors as well as non-genetic factors. Many of the concepts and methods in Fisher's paper became the foundations of quantitative genetics that we still use today.

The first quantitative trait loci (QTL) mapping was done by Karl Sax in 1923. He found that the weight of beans (*Phaseolus vulgaris*) followed a similar distribution to that of the pigmentation colors. The beans homozygous for color are about twice as heavy as the beans heterozygous for color. This observation suggested that either the Mendelian factor for color also affects weight as a quantitative factor, or there exists two tightly linked Mendelian factors that control the color and the weight of the beans, and that the effect on the weight is additive.

1.1.3 Genetic variation and markers

Mendel's law of segregation applies directly to alleles on different chromosomes. However, alleles on the same chromosome can be transmitted together as linkage groups and it is difficult to distinguish their individual effects. Recombination is the primary mechanism that separates them. In sexually reproducing diploid genomes, a pair of ho-

mologous chromosomes synapse followed by individual chromatids exchange segments of DNA in meiosis. The frequency of two genes being separated by a recombination event can be used to define their genetic distance, i.e. $m = -\frac{1}{2}ln(1-2r)$ by Haldane (Haldane, 1919), where r is the recombination frequency. It has been noticed that genetic distances do not uniformly distribute along the chromosome as the nucleotide distance, but instead have hotspots and cold spots (Jeffreys et al., 2005; Myers et al., 2005), and also varies considerably between genders (Kong et al., 2002). Recombination provides an important source of genetic variation and allows for evaluating the marginal effects of genes. Genetic markers that are experimentally accessible for capturing such variation are thus critical for mapping traits.

For most of the 20th century QTL studies have been greatly constrained by the lack of markers that can be densely spaced in the genome to capture the genetic variation. The development of DNA restriction fragment length polymorphism (RFLP) was the first method that substantially increased the resolution to DNA-level polymorphism. Eric Lander and David Botstein (Lander and Botstein, 1989) proposed statistical methods to dissect Mendelian factors in quantitative traits, which became the main stream approach in the following years.

In the last decade, technological advances made it possible to detect single nucleotide polymorphism (SNP) (see review Brookes, 1999), which is the most abundant form of genetic variation and offers a single base pair resolution. The International HapMap Project (The International HapMap 3 Consortium, 2010) is one of the key resources in defining a map of SNPs using nucleotide arrays. The project eventually genotyped 1.6 million SNPs in 1,184 individuals from eleven populations, focusing mostly on the common variants with allele frequency >5%. The 1000 Genomes Project was the first project to sequence a large number of individuals with a goal of cataloging genetic variation across populations. The pilot phase of the project (The 1000 Genomes Consortium, 2010) has identified 15 million SNPs, 1 million short insertion and deletions and 20,000 structural variants in 179 individuals from four populations. The most recent phase of the project (phase three) has identified 80 million SNPs in 2,523 individuals from 26 populations (unpublished). These projects have provided essential information for genetic mappings. Recently a num-

ber of projects aim to further improve cataloging genetic variation by sequencing a large number of individuals with particular focuses. This includes population wide sequencing project such as UK10K (http://www.uk10k.org/) for the British population, GoNL (http://www.nlgenome.nl/) for the Netherlands population, and SardiNIA (http://genome.sph.umich.edu/wiki/SardiNIA) for the Sardinian population, or disease focused projects such as the GoT2D for type 2 diabetes and the International Cancer Genome Project (International Cancer Genome Consortium, 2010, https://icgc.org/) for cancer.

1.2 Mapping quantitative traits

Heritable factors can be inferred from phenotypic distributions such as the frequencies of peas with different colors in Mendel's experiments. However, most traits do not have an intuitive phenotypic distribution as that of the pea color. The distributions can be very complex, particularly when a phenotype is controlled by multiple loci. In these cases, the marginal effect of individuals genes can hardly be detected or distinguished, and QTLs can not be discovered by modeling phenotype data only.

Using information provided by the molecular markers is an obvious way to resolve this puzzle. Although it is not possible to know the locations of the QTLs beforehand, with a dense marker map, some of the tested markers are likely to be in linkage disequilibrium with genuine QTL loci. These tagging markers can be mapped in a number of approaches, and quantitative methods have been developed to define the relationships between the markers and the traits.

1.2.1 Linkage analysis and its limitation

Linkage analysis aims to identify genetic factors influencing traits by analyzing the cosegregation of markers with the traits across generations in families. Based on this idea, linkage analysis has been tremendously successful in identifying Mendelian diseases. Some of the examples include the identification of multiple mutations in the CFTR gene causing cystic fibrosis (Tsui et al., 1985; Riordan et al., 1989), the disease

haplotypes in Huntington's disease (Gusella et al., 1983; MacDonald et al., 1992) etc. The susceptible variants are often rare, possibly shaped by negative selection, but for the method to work they need to be highly penetrant.

Linkage analysis has also been applied to common diseases and quantitative traits. For example, variants have been identified associated with inflammatory bowel disease (IBD) (reviewed in Mathew and Lewis, 2004), type I diabetes (Luo et al., 1995; Mein et al., 1998) and schizophrenia (Williams et al., 1999; Ekelund et al., 2000). However, the heritability accounted for by the identified variants is typically very modest, even when the heritability of the disease is much higher, e.g. IBD and schizophrenia. Clearly, the reported loci are only a fraction of the full picture of the genetic architecture of these diseases. When the genetic genetic architecture is complex, where the phenotype is determined by a collection of variants with low penetrance, often a very large number of families is needed to discover and differentiate these effects. For example the association of type 2 diabetes with the Pro12Ala variant in the peroxisome proliferative activated receptor-γ gene (PPARG), which has an effect size of 1.25 fold, could only be detected using linkage studies of over one million sib pairs (Altshuler et al., 2000). It is impractical to recruit enough families with several affected generations to obtain a sufficient number of informative meioses, especially given that human families tend to be small. It becomes even more challenging if the study disease has late onset. These results suggest that in contrast to Mendelian disease, where a limited number of high penetrance loci are responsible for the disease phenotypes, complex diseases have much more complex genetic architecture that the linkage analysis approach is not well powered to discover.

1.2.2 Population association analysis

Instead of using a linkage study design, a simple statistical association can be used, which merely states the co-occurrence of genotypes and phenotypes in a population. Such an association may exist due to the fact that a DNA segment that contains a variant affecting disease susceptibility can be inherited by many individuals that share a common ancestor who carries the factor. This approach has been extremely powerful

over the last 8 years, resulting in over 10,000 genotypes to phenotype associations (Wellcome Trust Case Control Consortium, 2007; The NHGRI GWAS Catalog, Welter et al., 2014).

1.2.2.1 Mapping disease variants with case control phenotypes

For many diseases, there is no clear quantitative measure indicating the disease status. The phenotype is then reduced to a binary form of whether an individual does or does not have the disease. In this scenario, a case control design is often used. Healthy and disease individuals, often on the order of thousands, are recruited to a study and genotyped for a large number of variants, on the order of hundreds of thousands to millions. The basic idea is to compare the genotype frequency of the markers between the cases and the controls. A highly divergent marker frequency would suggest a possible link between the marker and the disease status. For each variant, the number of individuals with AA, AB and BB genotypes can be counted for the healthy individuals (m_{0j}) and the disease individuals (m_{1j}) to form a contingency table as below.

Genotype	AA	AB	BB	Total
Case	m_{11}	m_{12}	m_{13}	$m_{1.}$
Control	m_{01}	m_{02}	m_{03}	$m_{0.}$
Total	$m_{.1}$	$m_{.2}$	$m_{.3}$	m

The association can be tested using a χ^2 test: $\chi^2 = \sum_{i=0}^{1}\sum_{j=1}^{3}\frac{(m_{ij}-E[m_{ij}])^2}{E[m_{ij}]}$ with two degrees of freedom, where $E[m_{ij}] = \frac{m_{i.}m_{.j}}{m}$. The effect of a genotype can then be estimated as an odds ratio $OR_{AA} = \frac{m_{13}/m_{11}}{m_{03}/m_{01}}$.

1.2.2.2 Mapping QTLs using a simple *t*-test

A variety of methods have been developed for QTL mapping (Leal, 1998; Balding et al., 2008). Here I introduce the widely used t–tests, the analysis of variance and more recently linear mixed models.

The marginal effect of substituting allele A with allele B can be evaluated by comparing the homozygous individuals AA and heterozygous individuals AB, assuming effects are normally distributed in each genotype group with same variance. Let μ_0 and μ_1 be the genuine means of the phenotypes. The test hypothesis can be formulated as $H_0 : \mu_0 = \mu_1$, and $H_1 : \mu_0 \neq \mu_1$. The test statistic is thus

$$t = \frac{m_1 - m_0}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_0})}},\ s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

where (m_0, m_1), (n_0, n_1), and (s_0, s_1) are the sample means, samples sizes and sample standard deviations of AA and AB respectively. H_0 is rejected if t exceeds a significant threshold, such as $\alpha = 5\%$ when $t > t_{(0.025)}$ for a two tailed test with n_0+n_1-2 degrees of freedom.

1.2.2.3 Mapping QTLs using linear regression models

More generally, population samples contain three genotypes (AA, AB, BB). The quantitative phenotype y can be modeled as resulting from the sum of genetic effects and environmental effects in a simple linear model

$$y_i = \mu + \beta x_i + \epsilon_i,\ i = 1...n,$$

where y_i is the phenotypic value of the ith individual; x_i is the genetic dosage of the ith individual, which is the allele count of one allele such as (0,1,2) for the number of B alleles in genotypes (AA, AB, BB). ϵ represents the random error in y that can not explained by x, which is assumed to be independently and identically distributed. To satisfy the assumption for the distribution of the error term, often phenotypic measurements need to be transformed, e.g. using log, square root or mapping to normal quantiles. This can also be extended to generalized linear models that allow for response variables that have a variety of error models.

A simple way to make inference about the parameters (β, σ^2) is to use the least

square approach, which gives

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y},\ \hat{\sigma}^2{=}\frac{1}{n-1}(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$$

where $\mathbf{y}^T = (y_1, y_2, ..., y_n)$ is a vector of phenotypes and $\mathbf{X}_j^T = (x_{j1}, x_{j2}, ..., x_{jn})$ is a vector of genotypes at variant j for each individual. β is often interpreted as the effect size, representing the contribution of a unit change in the genetic dosage encoded in x to the phenotype y.

This is equivalent to a single factor analysis of variance (ANOVA) of the genetic effect. The mean sum of squares within genotype groups SS_{within} reflects any other residual variation that is non-genetic. The difference between the total sum of squares (SS_{total}) and SS_{within}, $SS_{between}$, reflects the QTL genotypes effect on the phenotypes. The ratio between them $\frac{SS_{between}/(3-1)}{SS_{within}(n-1)}$ is an F value that can be used to test for the genetic effect. The statistical significance level can be computed by comparing to the F distribution with degrees of freedom 2 and $n-3$. If we let $\sigma_e$be the environmental variance and $\sigma_g$be the genetic variance, the heritability can be expressed as $h^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$, where σ_e can be estimated by $SS_{within}/(n-1)$ and σ_g^2 can be estimated by $(SS_{between} - SS_{within})/k$ where k is a factor adjusting for group size of three genotype groups ($k = 3/(\frac{1}{n_0} + \frac{1}{n_1} + \frac{1}{n_2})$). In case of comparing two genotype groups, $F = t^2$.

Many studies also use maximum likelihood approaches to estimate genetic effects. With the same linear model, the likelihood function is

$$L(\mu, \beta, \sigma) = \prod_i^n z(y_i - (\mu + \beta x_i), \sigma^2)$$

where z is the standard normal density function $z(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. The inference of the parameters is often done using the Expectation-Maximization algorithm (Moon, 1996). The likelihood of the full model with estimates $(\hat{\mu}, \hat{\beta}, \hat{\sigma})$ can be compared against a null model $(\hat{\mu_0}, 0, \hat{\sigma_0})$ where the genetic effect is removed by setting $\beta = 0$

to compute likelihood ratio statistic

$$LOD = -2Log(\frac{L(\hat{\mu}, \hat{\beta}, \hat{\sigma})}{L(\hat{\mu_0}, 0, \hat{\sigma_0})})$$

LOD has an asymptotic χ^2 distribution with one degree of freedom, which can be used to determine statistical significance.

1.2.2.4 Mapping QTLs using linear mixed models

Spurious associations can arise when study samples in association analysis have variable genetic relationships, in which case the ϵ_i are not independent. Such confounding factors of relatedness may not be known to the researcher from phenotypic data collection. To adjust for it, the linear mixed model approach has become a popular method of choice recently. These models typically use an additional random variable with a specific covariance structure to capture the genome wide sample relatedness (Kang et al., 2008; Zhang et al., 2010; Listgarten et al., 2012; Zhou and Stephens, 2012):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$$

where y is the phenotype vector, β is an unknown fixed effect for the candidate genetic marker, and u is an additional random effect reflecting the genetic effect due to relatedness. u is normally distributed with $u \sim N(0, K\sigma_g^2)$, where K is the kinship matrix, with each $k_{i,j}$ the correlation between the markers either genome wide (Kang et al., 2010) or a selected subset (Listgarten et al., 2012), of individual i and j. σ_g^2 is the unknown genetic variance. For statistical testing, similarly, a likelihood ratio statistic can be computed by comparing against a null model. This model successfully removes false positives due to sample structure. It can also help to refine genuine signals by controlling for the other genetic markers that are not the candidate locus being tested, such as using only markers on chromosomes except the one that the test mark locates in (Listgarten et al., 2012).

The main limitation of the linear mixed model approach is the computation cost,

which in the full model is of the order of MN^3 (Kang et al., 2008). There have been improvements on reducing the cost by making approximations. One approach is based on the assumption that the genetic effects of total markers by u is approximately shared between individual markers, thus the relationship matrix only needs to be built once instead of each time per marker. The data is then rotated by the eigen-decomposition of the relationship matrix to remove the structure (Listgarten et al., 2012). Another approach is based on the observation that a relatively small number of independent markers can be selected to capture the information about relatedness. A careful selection of markers could dramatically reduce the size of the relationship matrix and allows for exact analysis in each test (Listgarten et al., 2012).

1.2.3 Multiple testing correction

In an association scan, a collection of statistical tests is typically evaluated for a large number of markers. While there are good reasons for doing so, such as one wishes to allow as many genetic causes as possible, this leads to a major issue in the greatly increased probability of declaring false positives. Typically a nominal $p = 0.05$ is used to claim an effect is statistically significant. This means that the probability of rejecting null hypothesis is 5% by chance. However, in cases where a data set is used to test for many hypothesis, the probability of reaching $p = 0.05$ is substantially elevated by chance. For example, in 100 tests, the probability of observing at least one test significant at 5% level is

$$Pr(\min_i p_i \leqslant 0.05) = 1 - Pr(all\ p_i > 0.05) = 1 - (1 - 0.05)^{100} = 0.99$$

which means it is almost guaranteed to have at lease one nominally significant association.

Methods have been developed to resolve this problem by adjusting the threshold when multiple tests are performed. The minimum p value distribution, which is substantially skewed to low p values, is used as the p value distribution under the null hypothesis instead of the individual p value distribution, which is uniform. The corresponding error rate is often referred to as the family wise error rate (FWER).

The Bonferroni correction is a simple method to control for $FWER < \alpha$ by using a threshold of $p < \frac{\alpha}{m}$, where m is the number of tests.

Bonferroni correction can be too strict in many cases. A more liberal approach is to control for a false discovery rate, where the significance is declared while accepting a fraction of false positives. One popular method is the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), which relies on the assumption that p values under the null model are uniformly distributed. A false discovery rate can thus be calculated by comparing the observed value against its percentile rank: $p < \frac{k}{m}q$ to declare k signals out of m tests at a false discovery rate q.

More recently, a q value approach (Storey and Tibshirani, 2003) was developed and frequently used in many studies (Degner et al., 2012; Maurano et al., 2012; McVicker et al., 2013). It further calibrates the balance between the fraction of the declared significances and the false positives in an automated way. It estimates the proportion of tests π_0 that are drawn from the null by fitting a cubic spline to the p value distribution and taking the frequency of the p values at the end of the distribution, which reflects the proportions of nulls when there is no association. This π_0 can then be used to calibrate a false discovery rate at any p value level.

The distribution of p values under the null hypothesis can also be empirically estimated using permutations. This is normally conducted by random assigning phenotypes or genotypes to individuals. The nominal p values from the original test can be compared to the p values from the permutations to establish an FDR level.

1.2.4 Statistical power in genetic associations

The statistical power to detect associations between genotypes and phenotypes depends on a number of factors. Situations where variants have small effects are particularly hard to map. The linkage strength between a marker and a genuine QTL also adds to the complexity. Below I discuss how these factors relate to each other using the simplest t-test model.

Assume that we want to seek for results controlling for a type I error rate α and a

type II error rate β, then

$$1 - \beta = Prob(t_1 > z_{\alpha/2}) = 1 - \Phi(z_\alpha - t)$$

where z_α is the critical value for the confidence level $1 - \alpha$under the null hypothesis $t_1 = 0;\Phi$ is the standard normal cumulative density function. If we assume the ratio of AA:AB:BB is 1:2:1, a QTL is linked with the tested marker at a recombination rate r, and the genuine additive effect is a, then the difference between AA and BB is $m_2 - m_0 = (1 - 2r)2a$, where (m_0, m_2) are the phenotypic sample means of AA and BB. The t statistics is calculated as

$$t = \frac{m_2 - m_0}{\sqrt{s^2(\frac{4}{n} + \frac{4}{n})}} = \frac{(1 - 2r)2a}{\sqrt{8s^2/n}}$$

Replacing t with $z_{\alpha/2} + z_\beta$,

$$n = 8[\frac{z_{\alpha/2} + z_\beta}{(1 - 2r)2a/s}]^2$$

We can see that QTL can be detected with small n if the effect size a is large, the linkage r between the marker and the QTL is strong and the residual noise s is small. Note that the QTL effect is only mediated via the marker locus that is linked to the causative variant, thus the real effect can be under estimated, and it is not easily distinguishable between a strong effect via weak linkage and a weak effect via strong linkage.

So far the most reliable way to validate a discovery is to replicate the result in an independent sample cohort. The general principle of choosing the replicate setting is to repeat the initial experimental design as closely as possible, with samples drawn from the same population and phenotypically ascertained using the same procedure. The position of the associated loci in the replication cohort must be identical to the original position or in strong linkage disequilibrium, with an effect in similar order and in same direction. One caveat is that the effect in the initial association can be over estimated due to winner's curse (Zollner and Pritchard, 2007). On the other

hand, fewer tests are conducted in replication, reducing the multiple testing burden. Estimates from multiple replicates can regress towards the genuine mean effect.

1.2.5 GWAS results and interpretation

The Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007) performed the first large association study by comparing disease individuals and healthy individuals (case-control design) for 7 diseases. Some of the early successes using the GWAS approach include the discovery of TNFSF15 as susceptibility gene to the Crohn's disease (Yamazaki et al., 2005) and TCF2 (or HNF1B) for type 2 diabetes and prostate cancer (Gudmundsson et al., 2007). Variants are also found in genes that can be targeted by drugs, such as PPARG and KCNJ11 associated with type 2 diabetes, and IL12B associated with psoriasis, targeted by thiazolidinediones, sulfonylureas and anti-p40 antibodies (Krueger et al., 2007; Manolio et al., 2008). These early results were followed by an explosive growth of studies with more study individuals and using more dense genetic markers. Recently, the International IBD Genetics Consortium (IIBDGC) (http://www.ibdgenetics.org/) reported the discovery of 163 loci associated with Crohn's disease using a very large study cohort consisting of over 75,000 individuals (Jostins et al., 2012). Many of the loci reported in this study are implicated in other immune-mediated disorders, e.g. ankylosing spondylitis and psoriasis.

In some cases, the causal relationship is plausible, such as the IFIH1 gene identified as associated with diabetes (Nejentsev et al., 2009). The gene is known to play a role in antiviral infection and there is strong link between type 1 diabetes and viral infection(Nejentsev et al., 2009). However, more often there are cases where the functional relevance is not obvious. Variants reported in different studies sometimes reveal unexpected connections, e.g. CDKN2A is reported to be associated with Coronary disease, type 2 diabetes, and invasive melanoma (Kamb et al., 1994; Helgadottir et al., 2007; Scott et al., 2007).

One important observation in genome wide association studies is that the odds ratio for associated variants is modest, typically between 1 to 1.5 (Hindorff et al., 2009).

For example, a recent large scale genome wide association study on type 2 diabetes in more than 150,000 individuals revealed more than 70 loci but only explain 11% of T2D heritability(Morris et al., 2012). Similarly, a large scale study on Crohn's disease showed that the heritability is only 23% (Franke et al., 2010). The reasons are two fold. First, it is possible that the variants identified by the genome wide association study are only a small subset of the variants that contribute to the disease etiology. Due to the statistical power and the winner's curse, the rest is not sufficiently powered to be discovered, particularly the ones with low allele frequency, e.g. MAF<1%. Second, the genetic effect of a DNA variant must propagate through multiple levels of cellular networks, regulated by other mechanisms such as epigenetics or by environments, which substantially reduces the initial effect, manifesting a weak effect at higher level that can result from initial strong effects at cellular level.

Albeit the challenges and limitations, the GWAS results nevertheless highlight informative clues on the underlying biology. To seek further understanding, one has to investigate deeper into tissues and cells, these reasons have motivated studies into mapping molecular phenotypes measured in a cell, where most statistical methods are also applicable.

1.3 The promise of cellular phenotypes

1.3.1 Moving towards cellular phenotypes

Cellular processes are more directly subject to genetic regulation. For that reason, the effect size, defined as the magnitude of change in the downstream measurement by a change in the genetic allele, could be much higher than individual level traits.

Another important advantage is that cellular phenotypes can be linked to interpretable cellular products. The regulation process can be seen as a generative process, starting from decoding the information stored in DNA to transcribing into RNA and then to translating into proteins. The measurement of the product abundance at each step could reveal the mechanistic process with direct relational context. In practice, it is not yet technically possible to capture all these types of information simultane-

ously, but it is already feasible to measure each step separately using various molecular assays and integrate the measurements later in computational analysis.

1.3.2 The measurement of cellular phenotypes

The first major leap in large scale measurement of cellular phenotypes is perhaps the microarray. It is based on the same idea as Southern Blot (Southern, 1992) that DNA fragment can be hybridized to known complementary DNA sequences, called probes. This can be used to measure gene expression levels, where mRNA is reverse transcribed into cDNA, which can then be hybridized to a microarray. The method allows simultaneous quantitation of a large number of probes, designed to target a number of genes. The first study using microarray profiling gene expression was published in 1995 (Schena et al., 1995).

In 2005, the birth of next generation sequencing started a new era for genomic assays. It has further revolutionized the sequencing of DNA using an idea of sequencing by synthesis for a large amount of short DNA fragments in a massively parallel way(Bentley et al., 2008). The price has dropped exponentially as a result, from $10M in 2005 to $4000 in 2014 per human genome at 30x coverage (NHGRI, www.genome.gov/sequencingcosts). It is gradually making investigating genetics at genome wide scale for a large group of individuals practically feasible.

Next generation sequencing technology also gives rise to a large variety of assays that are designed to measure other molecules. The basic idea is to transform the desired molecular information into a collection of DNA sequences, which can be sequenced. During the past few years, a rich collection of methods have been developed. For example, for RNA transcription related information, RNA-seq (transcript abundance, Chu and Corey, 2012) and GRO-Seq (binding sites of active Pol II, Core et al., 2008); For translation, Ribo-Seq (ribosome profiling, Ingolia, 2014) etc have been developed. For DNA Methylation, Bisulfite Sequencing (BS-Seq, Krueger et al., 2012) and MeDIP-Seq (Taiwo et al., 2012) are widely used. For DNA-Protein interactions, there are DNase-Seq, FAIRE-Seq and ChIP-Seq (See review Furey, 2012). A recent refinement of ChIP-seq, ChIP-exo, is able to identify the exact bases that are bound

by a factor (Rhee and Pugh, 2011). Chromosome conformation can be measured by assays such as Hi-C/3C-Seq and more recently 5C, which relies on the cross linked DNA generated due to interactions between two factors (Dostie and Dekker, 2007; Simonis et al., 2009; Lieberman-Aiden et al., 2009). There are also assays designed to measure special sequence elements, such as Tn-Seq for transposon sequencing.

The resulting sequences from these assays can be aligned to a reference sequence to reveal the information about where the event has occurred and how much of target products exist in the starting material. Chapter 2 of this thesis uses ChIP-seq technology in particular for measuring bindings of the CCCTC (CTCF) binding factor. In detail, it works by extracting segments involved in protein-DNA interactions using Chromatin Immunoprecipitation(ChIP) followed by sequencing. When protein-DNA interaction occurs in a cell, binding proteins and DNA segments are temporarily bonded as a complex. Such structure can be chemically strengthened using cross linking agent, after which the long DNA molecules are then shared into ~500bp fragments by sonication. This produces a mixture of DNA fragments, within which some are bonded by proteins. The ones of interest are then selectively immunoprecipitated from cell debris using specific antibodies, such as anti-CTCF in the case of chapter 2. The target molecule is thus enriched and purified. Once this material is obtained, the associated DNA fragments can be extracted out and sent for sequencing. The initial locations of the protein-DNA interactions can then be determined by aligning these sequences back to the reference genome. The quantity of the fragments corresponds to the number of the molecules in the starting material, representing the intensity of the binding. One caveat is that there is a number of sources of technical variation involved in the data production (Taub et al., 2010). For example, non specific fragments may remain in the purified material, which then become background for the real binding sites. Computational methods have been developed to differentiate signals from background.

The fast development of cellular assays has opened the door to obtain cellular information in an economical, genome wide, and simultaneous way. This has allowed to investigate the genetic landscape of molecular traits such as gene expression, transcription factor binding, histone modification etc. The relations or dependencies between

these molecular events can be examined in a scale that has never been reached before. QTL approaches can be applied to discover genetic loci that play a role in regulation in various levels and aspects of the molecular processes in a cell.

1.3.3 Latent variables in analyzing high dimensional genotypes and cellular phenotypes

In association mapping with disease traits, tens of thousands to tens of millions genotypes are assayed. In QTL mapping of cellular phenotypes, in addition to a large number of genotypes, a high dimensional phenotype is also measured, e.g. by microarray or by next generation sequencing assays. The measured phenotypic variation can come from sources such as cellular fluctuations (Liebermeister, 2002; Dueck et al., 2005; Gibson, 2008), regulation of gene expression (Sanguinetti et al., 2006; Pournara and Wernisch, 2007), and environmental conditions (Hastie et al., 2000), many of which are confounding factors that need to be accounted for to prevent loss of power in discovering true signals and also false discovery of spurious signals (Leek and Storey, 2007; Hyun et al., 2008).

One way of disentangling the mixture in a high dimensional dataset is to use dimension reduction techniques to identify key components that reflect the data structure. On the one hand, these components can be used to reflect the relationships between samples learned from the measured dataset, thus become useful indicators when independent sampling is assumed. On the other hand, they can help identify sources that are influential to a large number of traits, which often come from a non-interesting source, such as technical batches. Data for some of the factors that affect transcript levels may have been collected by researchers, such as age, experimental batch, etc., the inclusion of which as covariates in association models have shown to improve QTL discoveries (Emilsson et al., 2008). However, perhaps more prevalently, the confounding factors are hidden to researchers. In this case computational methods can help identify them, which can be considered alongside known covariates in association mappings (Leek and Storey, 2007; Hyun et al., 2008; Stegle et al., 2010; Nica et al., 2011; Fusi et al., 2012). At a smaller scale, such as genes in a pathway, such component

can themselves become phenotypes that reflect the commonality of the traits that are functionally linked. This thesis explores this in Chapter 3 in the context of associating gene expression to ageing phenotypes.

Among the dimension reduction techniques, principal component analysis (PCA) is perhaps the most widely used method. In PCA the original data is converted by an orthogonal projection onto a lower dimensional linear space known as principal space. The corresponding dimensions, known as principal components, are learned from the data by either maximizing the variance of the projected data (Hotelling, 1933) or equivalently minimizing the averaged projection costs defined as the mean square error between the projections and the initial data points (Pearson, 1901). Consider a dataset $\{\mathbf{x}_n\},\ n = 1, ..., N$ where $\mathbf{x}_n = \{x_{nm}\},$ m=$\{1,...\text{M}\}$ with N observations each of M dimensions. PCA attempts to project the data onto a space with dimension $D < M$. The dimension of the new linear space can be defined by a unit vector $\mathbf{u}$ with constraint $\mathbf{u}^T\mathbf{u} = 1$. The variance of the projected data is thus given by $\mathbf{V} = \mathbf{u}^T\mathbf{S}\mathbf{u}$ where $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$. Maximize $\mathbf{V}$ with the constraint on $\mathbf{u}$ gives a quantity λ that satisfies $\mathbf{S}\mathbf{u} = \lambda\mathbf{u}$, where $\mathbf{u}$ and λ are the eigenvector and eigenvalue of $\mathbf{S}$. This process can be repeated to obtain D principal components $\{\mathbf{u}_1, ..., \mathbf{u}_d\}$ with corresponding eigenvalues $\{\lambda_1, ..., \lambda_m\}$ with $\lambda_1 > \lambda_2 > ... > \lambda_d$.

As a non parametric method, PCA has an advantage of not requiring model assumptions. Other advantages also include fast computational speed, and very easy visualization for separating samples based on their high dimensional measurements, e.g. Novembre et al. (2008) showed the first two principal components learned from one million genotypes correctly separated European populations into geographic groups. The PC projection can also be directly linked to the genealogical history of samples (McVean, 2009), although this may not be unique as multiple processes such as isolation, migration and admixture can give similar projections. In disease mappings, because of this property, PCs are useful to control for population stratification, where they can be included as fixed covariates in association models (Price et al., 2006; Novembre and Stephens, 2008).

Alternative to the linear projection, PCA can also be expressed as a probabilistic solution for latent variables using maximum likelihood, known as the probabilistic PCA

(Tipping and Bishop, 1997). Consider $\mathbf{z}$ latent variables corresponding to the principal components with a prior probability $p(\mathbf{z}) = N(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix, the conditional probability of the observations is $p(\mathbf{x} \mid \mathbf{z}) = N(x \mid \mathbf{Wz} + \mu, \sigma^2\mathbf{I})$, where the columns of $\mathbf{W}$ define a D dimensional linear subspace. The observations in $\mathbf{x}$ is thus reconstructed from a mapping between a space spanned by $\mathbf{W}$ to the original data space by $\mathbf{x}$, with additional Gaussian noise with a variance of σ^2. Different from the conventional PCA, the numerical solutions of $\mathbf{W}$ such as via the EM algorithm do not guarantee that the columns of $\mathbf{W}$ are orthogonal to each other.

Probabilistic PCA has some advantages over the conventional PCA. This includes more efficient inference using the EM algorithm, having a likelihood function that can be readily used for comparison with other models, automatic identification of the dimension of the subspace due to the Bayesian treatment etc (Bishop, 2006). A closely related method to the probabilistic PCA is factor analysis (Basilevsky, 1994; Tipping and Bishop, 1997). The only difference is that its covariance structure is defined as $\mathbf{\Psi}$, an $M \times M$ matrix, instead of an isotropic $\sigma^2\mathbf{I}$, where $\mathbf{\Psi}$ captures the independent variance associated with each coordinate. This feature has shown to be particularly useful in capturing natural correlation between variables, such as expression levels of genes that are functionally linked. The PEER package (Stegle et al., 2010) provides software for both conventional PCA and factor analysis for high dimensional genomic data.

1.3.4 The genetics of gene expression

As an important product of DNA coding, gene expression is technically feasible to measure and has thus attracted a large amount of research focus. Many studies have looked for expression QTLs (eQTLs), aiming to link genetic variation to expression levels of gene products. Genetic regulation of gene expression levels has been found to underlie phenotypes from human diseases (see reviews Kleinjan and van Heyningen, 2005; Wray and Wray, 2007) to the morphology of Darwin's finches (Abzhanov et al., 2004).

Genetics of gene expression variation Gene expression variation can result from a number of factors, including environmental effects, epigenetic effects, biological random fluctuations, and genetic effects. How much of phenotypic variation can be explained by genetics is one of the core questions that need to be addressed. Studies have shown that a high proportion of gene expression levels (over 40%) are heritable across individuals (Petretto et al., 2006; Stranger et al., 2007; Dixon et al., 2007; Göring et al., 2007; Price et al., 2011; Grundberg et al., 2012). The heritability varies but is mostly greater than 10%, which is much larger than that from a typical GWAS study for whole organism phenotypes (see review Skelly et al., 2009).

A recent study on over four hundred twins found 8,329 out of the 13,970 genes in the investigation have shown one or several QTLs (Lappalainen et al., 2013). This suggests that the expression levels of a large majority of the genes are under genetic control. An additional level of evidence supporting this can be seen when comparing the expression levels of the two alleles of a gene within heterozygous individuals, or the allele specific expression (Morley et al., 2004). This study has found in humans around 6.5% of sites per individual show allele specific expression, largely consistent with the eQTLs. As the sample size and the accuracy of measurement increases, more eQTLs are likely to be discovered, including ones with relative weak effects (Cheung et al., 2010; Grundberg et al., 2012). The current data suggest that variation in the expression of genes is substantially linked to the genetic background.

The eQTLs discovered so far are enriched in regions close to the transcription start site (TSS) and the transcription end site (TES), areas known to play a role in the regulation of gene expression, mostly via transcription factor binding or methylation modifications (Veyrieras et al., 2008; Dimas et al., 2009; Stranger et al., 2012). Many eQTLs are found within the promoter binding motif, directly perturbing the binding in the interface, making them very likely to be causal. These effects are presumably mediated via changing the binding affinity of the promoter complex, which subsequently affects the efficiency of gene transcription. This may also cause differential usage of promoters, which is known to be an important source of variation in gene expression (Forrest et al., 2014). Notably, 16% of disease GWAS variants are eQTLs (Lappalainen et al., 2013), evidence supporting an effect route from DNA to

gene expression then to disease traits.

Genetic regulation in *cis* and *trans* Genetic effects on gene expression and other molecular phenotypes within a genomic location are often categorized into in *cis* and in *trans*. These term were initially introduced by Haldane (Needham, 1942) to describe different allele configurations in heterozygous individuals, with a meaning actually more similar to linkage disequilibrium. The terminology was later used by Lewis (Lewis, 1945) to describe whether two mutations are in the same gene.

In the eQTL literature, *cis* and *trans* are typically defined more based on the distance between the associated variants to the target genes. Genetic regions proximal to the target genes are referred to as *cis* regions while the ones at a different chromosome or far from the target genes are referred to as *trans* regions. This may be a reasonable classification as a large proportion of eQTLs are indeed close (<100kb) to the transcription start sites (TSS) of genes (**?**), representing an important type of *cis* elements. It however can also be problematic, for example the distance thresholds used for differentiating *cis* and *trans* is arbitrary in different studies, from a few hundred base pairs to one or two megabases.

Notably, the differentiation between the *cis* and *trans* effect can also be based on the principle that *cis* elements are allele specific, while *trans* element can act on both of the target alleles. One study design is to compare the ratio of transcription between the two alleles in a hybrid offspring and the gene expression levels between the two parents (Wittkopp et al., 2004). A consistent ratio would suggest a *cis* effect driven by the target gene while otherwise it suggests a *trans* effect driven by other factors somewhere else in the genome or epigenetics effects.

Localizing *cis* and *trans* effect elements involves hugely different levels of technical challenges. The search space for a *trans* association is the product of the number of genetic markers and the number of expression traits, which is several orders of magnitudes greater than that for a *cis* scan. As a result, *a trans* effect with a similar effect size as a *cis* effect is much harder to detect because of the multiple testing penalty. A *trans* scan also involves correcting for more confounding factors that further weakens the signal. Indeed, *trans* eQTLs discovered so far are only a small minority

(Stranger et al., 2007; Small et al., 2011) in human studies.

1.3.5 The genetics of transcription factor binding

Transcription factor binding variation is one of the primary mechanisms by which gene expression is modulated. Key questions include what is the variability of transcription factor binding, what drives it, and how does it affect variation of gene expression levels. Studies have largely taken one of two approaches: 1) investigate specific regions at which regulatory events occur to build transcription factor binding maps, and associate genetic sequence variation within the binding sites to the binding variation; 2) consider binding variation as a quantitative trait and apply QTL mapping.

Technically, transcription factor binding can be measured using ChIP-seq genome wide, which does not require prior knowledge of the binding sequence. The sequence reads from a ChIP-seq experiment can be aligned to the reference genome to recover where the binding events have occurred and how strong they are. This normally involves a computational analysis called peak calling, which essentially identifies regions with a higher density of reads compared to that in the background based on estimations using various models (see reviews Laajala et al., 2009; Park, 2009). Using the reads mapped at the identified binding peaks, algorithms have been developed to infer short sequence patterns, called motifs, with a length normally less than 20bp, predicted to be the binding interface between the transcription factor and the DNA nucleotide (Tompa et al., 2005; Elnitski et al., 2006).

Transcription factor binding variation Studies on binding variation between species suggest that many binding events are species specific, with large divergence between species. For example, Boyer et al. (2005) and Kunarso et al. (2010) showed that the binding of two key regulatory proteins (OCT4 and NANOG) in human and mouse embryonic stem cells show dramatic divergence. Such divergence was also seen in hepatocytes when comparing transcription factor binding profiles between human and mouse (FOXA2, HNF1A, HNF4A and HNF6, Odom et al., 2007). The binding profile are substantially diverged in closely related yeast (Ste12 and Tec1, Borneman

et al., 2007) and fungi (MCM1, Tuch et al., 2008). A recent study comparing two transcription factors in five vertebrates reconfirmed the pattern shown in the previous studies (Schmidt et al., 2010), revealing that individual binding events are gained and lost rapidly during evolutionary time, although the conservation level varies largely between different transcription factors, suggesting different evolutionary constraint.

There is also substantial binding variation between individuals within species. Zheng et al. (2010) found 30% of sites of STE12 show binding variation in a group of yeast segregants from two divergent parents. Kasowski et al. (2010) profiled NFκb and Pol II in a small group of ten humans and showed that 25% and 7.5% respectively of sites vary between individuals. A subset of the binding variation correlates with downstream gene expression. This suggests that many differences in individuals and species are at the level of transcription factor binding, which plays a strong role in species diversity and gene regulation.

Genetics of transcription factor binding variation It is of great interest to understand what gives rises to the variation in transcription factor binding. Genetic factors and environmental effects can both play a role. Recent studies have increasingly shown that heritable genetic effects are responsible for a large component of the transcription binding variation. A clever experiment by Wilson et al. (2008) provided convincing evidence. The study used an aneuploid mouse strain carrying a human chromosome 21, and asked whether transcription factor binding on chromosome 21 is driven by human sequence or by the mouse nuclear environment. The results showed that transcription factor binding on the human chromosome is largely recapitulated, supporting the hypothesis that transcription factor binding is mostly governed by genetic sequence.

Associating genetic variation with binding variation in yeast has showed that *cis* regulation plays the primary role (Zheng et al., 2010). The linked genetic variants tend to reside within the binding motif of the target protein or related cofactors. This suggests that genetics affects transcription factor binding by affecting the binding affinity at the protein-DNA interface. The variants that affect sequence motif that subsequently affect binding affinity correlate with the binding signals (Kasowski

et al., 2010; McDaniell et al., 2010). It is also common that some binding events of transcription factors are correlated with mutations near the binding motif (Kasowski et al., 2010; McDaniell et al., 2010). Notably, binding variation also depends on the accessibility of the DNA in chromatin configuration, supported by the findings that sequence variation that affects DNase I sensitivity sites, nucleosome positioning and DNA methylation also affect transcription factor binding (Segal and Widom, 2009).

Validation of the function of transcription factor binding A variety of computational methods have been developed to infer the functions of QTL variants, mostly by summing evidence from published functional data sets as well as sequencing conservation (McLaren et al., 2010; Kircher et al., 2014). Eventually, an experimental validation such as by gene knock-down or nuclei base editing (Cong et al., 2013; Hwang et al., 2013) will be required to confirm the predictions. For example, Cusanovich et al. (2014) investigated differential transcription factor binding by knocking down 59 transcription factors in one HapMap lymphoblastoid cell line. The results show that most transcription factor changes only exert weak impact on the expression levels of genes within a 10kb window, and the ones that cause large changes tend to be located at transcription factor binding clusters, or at sites with high binding affinity or at enhancer regions. In a related study, the FANTOM consortium (The Fantom Consortium, 2014) knocked down 52 transcription factors in an acute monocytic leukemia-derived cell line (THP-1) throughout a time course of growth arrest and differentiation (Suzuki et al., 2009), revealing complex roles of transcription factors in the regulatory network, with no single transcription factor driving the differentiation process. These studies have identified a small number of functional transcription factors or core regulators, a perturbation of which cause immediate downstream gene expression changes.

In general, the connection between the transcription factor network and gene expression appears to be complex with individual effects being relatively weak. It is possible that the perturbation of a single transcription factor can be compensated by other factors in the same biological process. Sophisticated system biology approaches may help reveal the network relationships and their impact on the gene expression.

It is also noted that, although there has been a large volume of studies on transcription factor binding, primarily driven by technological advances such as ChIP-seq, it is still not possible to profile all transcription factors in a cell. High quality antibodies are still not available for all factors due to technical limitations. The scope of current studies is largely affected by this technical limitation to focus on a small number of transcription factors whose measurement is technically robust. A much bigger picture of the binding landscape is yet to be revealed.

1.3.6 The genetics of other epigenetic variation

The DNA molecule is physically organized into a three dimensional structure of chromatin. The scaffold of the structure that DNA coils around is made by protein complexes called nucleosomes that are composed of histone protein octamers. Regulatory information is conveyed by the positions of nucleosomes and the modification of histone proteins, the tails of which can be covalently modified by methylation or acetylation (Campos and Reinberg, 2009; Segal and Widom, 2009). Such modifications have been shown to correlate with downstream functions. Covalent modification can also occur on nucleotide with methyl groups added to cytosine. These modifications, which interact closely with DNA nucleotide itself and play important roles in the readout of DNA information, are generally termed as epigenetics.

Epigenetic elements involved in organizing chromatin structure The architecture of chromatin is not completely understood. Studies have shown that there exist areas that are attached to the nuclear lamina forming a particular spatial organization. These lamina associated domains are surrounded by CpG islands and insulators such as the CCCTC binding factor, and are associated with low gene expression (Guelen et al., 2008). These results suggest a functional impact of chromatin structure by delineating broad active or recessive environments for the readout of DNA information by transcription. A related study applied Hi-C technology to identify higher order chromatin interactions genome wide in human and mouse embryonic stem cells. It identified “topological domains” that are particularly involved in the interactions,

and these domains are correlated with insulator binding protein CTCF, housekeeping genes, tRNA genes and short interspersed element (SINE) retrotransposons (Dixon et al., 2012). The active chromatin areas also correlate with open chromatin structure, with DNA in linear form and not wrapped around nucleosomes. These areas can be identified by using restriction enzyme DNase I, as only the open chromatin areas are exposed to excision sites that can be digested.

Genetic factors in epigenetic variation Epigenetic variation can result from genetic or non-genetic reasons. It is known that epigenetic modification helps to store the memory of the environmental exposures. A key question is to what extent epigenetic variation between individuals are due to genetic reasons? A related study (McDaniell et al., 2010) found that DNaseI foot print is highly heritable using six samples from a family of European ancestry. Another more recent study performed DNase I hypersensitivity site mapping in 70 HapMap cell lines of Yoruba ancestry, and identified a large number of genetic variants that are associated with the level of chromatin accessibility (dsQTLs). It estimates that over 50% of eQTLs are dsQTLs, with their effects mediated through chromatin accessibility. Based on the same set of cell lines, Bell et al. (2011) discovered methylation levels of 180 CpG-sites in 173 genes associated with *cis* QTL variants (10% FDR). Another study (Zhang et al., 2014) discovered that cytosine modifications at CpG sites are primarily driven by *cis* QTLs using over one hundred HapMap cell lines of European and African origin. A subset of these modifications colocalize with transcription factors to enhance or repress gene expression, often associated with changes in chromosome accessibility. These studies have established the regulatory connections between genetic variation and epigenetic variation (similar results are seen in McVicker et al., 2013).

Non-genetic factors in epigenetic variation It has also been seen that there exist substantial non-genetic causes for epigenetic variation. Environmental effects can also cause methylation variation thus in general the direction of causality is unknown. A change in methylation can either be the result of a genetic effect changing it to the current status, or an environmental exposure that is stored in a form of epigenetic

modifications. The monozygotic twins with identical genetic background can help resolve this, as MZ twins have identical genetic background (Bell and Spector, 2012). In the aging context, Bell and Spector (2012) showed that differential methylation is associated with age and age related phenotypes in a twins cohort, but highlighted that a subset of these can be mediated by genetic reasons. Another study (Rakyan et al., 2011) investigated monozygotic twins pairs that are discordant for childhood-onset type 1 diabetes (T1D) identified methylation sites that are associated with the disease.

1.3.7 Tissue and environment effect in QTL mapping

One important caveat in QTL mapping is cell type and cell state. Studies on multiple tissues have shown that although there is a modest degree of sharing, quite often the regulatory effect of an eQTL is private to a specific tissue. It is therefore crucial to search for QTL in the correct tissue, i.e. in a tissue that is relevant to the disease of interest. This may not always be straightforward, as in some cases the regulatory effect is not in the tissue where a trait is manifested. Using a wrong tissue could be misleading. Cells with similar differentiation lineages have increased eQTL sharing relative to developmental distant tissues. However, a significant fraction of *cis*-eQTLs are cell type specific. This argues that variation that primarily affects late developmental processes may achieve sufficient power to be discovered by a *cis* scan (see review Gaffney et al., 2012). A number of projects have started to look into the landscape of gene expression as well as regulatory element signals in multiple cell types (**?**; The Fantom Consortium, 2014). This is still very challenging, as many tissues are not experimentally accessible, or are financially expensive when studied on a large scale, which can be necessary to detect weak effects or intra individual effects. The Human Induced Pluripotent Stem Cells Initiative (HipSci, http://www.hipsci.org/) is one of the first projects to systematically investigate genetics, epigenetic, proteomics and cell biology in induced stem cells and the differentiated daughter cells from them. With induced stem cells, HipSci is able to access tissues that are normally not very accessible from normal sample biopsy procedure, such as neurons, by differentiating

stem cells into the target cells. Genetic mappings for a variety of tissues obtained in a number of differentiation stages may be powerful in revealing some of the key biological insights, such as how genetics regulates tissue differentiation. This could be the first step towards understanding this important process, and ideally one should measure such process *in vivo*.

Additional to cell type, cell state also contributes to the variation of molecular phenotypes. It is known that the transcription profile changes dramatically when a cell is at different stages of its life cycle (Marguerat et al., 2012). Most current studies are conducted in cells in quiescent state, which may not be the state relevant to the trait. Phenotypes are not necessarily present in the quiescent state, and the ones of interest can be hidden in this system. Applying assays that are targeting the correct cell state will be important in reveal the genuine regulatory architecture.

Environmental exposure is also an important source of variation. In the absence of accurate measurement, environmental factors will cause loss of power due to increased stochasticity. Most environmental exposures are hard to measure. The number of study cohorts with well annotated environmental measures, usually obtained from questionnaires, is very limited, and even if there is, it is not certain that the relevant quantities are measured.

1.3.8 Resolving the causative relationship

Cellular molecules work together in a system to achieve a biological function. Genes responsible for a particular biological function can be grouped together as a pathway, e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) catalogs some common pathways (Kanehisa et al., 2004). It is of interest to know the causative relationships among the molecules, which gives the knowledge of how a function is achieved. It is possible to do this from experiments by perturbing different combinations of various molecule levels. The relationship can however also be estimated numerically, for example, looking at conditional probabilities in different regulation configurations (e.g. Schadt et al., 2005). In the genomic context, one important characteristic is that genetic variation is generally fixed in an individual's life time, with the exception of

somatic mutations. Thus it provides an important anchor to resolve such relationships (Lawlor et al., 2008), which is tremendously interesting in understanding its role in the network.

1.4 Overview of the remainder of this thesis

This thesis tackles a number of problems of mapping cellular traits as well as developing new methods for measuring cellular phenotypes. The remaining of the thesis is organized in four chapters

- Chapter 2 describes a first well powered systematic QTL study of a primary transcription factor CCCTC binding factor. This work is published in PLoS Genetics (Ding et al., 2014).
- Chapter 3 describes a method to substantially increase power in gene expression association to ageing. This work is accepted in G3 subject to minor revisions (joint first author).
- Chapter 4 describes a novel approach to measure telomere length using existing genome or exome sequencing data in a large scale. This work is published in Nucleic Acid Research (Ding et al., 2014, first author). The method has been applied to a melanoma study, which discovers several mutations in the Protection of Telomeres 1 (POT1) gene that are disease susceptible (Robles-Espinoza et al., 2014).
- Chapter 5 provides a conclusion, drawing together materials from the previous chapters and discusses future directions.