Genetic Epidemiology - How to Quantify, Localize and Identify Genetic Influences on Human Traits
Introduction

Individual differences in psychological or behavioral traits can be explained by a combination of genetic and environmental differences between individuals. When a trait is said to be highly genetic, this means that a large proportion of the variance in the trait is explained by genetic factors, i.e. the effect of one or many genes that each have their influence on the expression of the trait. Environmental factors can range from intrauterine environment to the influence of the family environment, school, friends and many other unidentified non-genetic factors. In this chapter we will provide an overview of methods used to model the contribution of genes and environment to variance in a trait or a set of traits, and to localize and identify the regions of the genome that may be involved. The area of research that focuses on quantifying genetic effects is called behavior genetics or genetic epidemiology. Genes can be localized and identified with genetic linkage and association methods. Finally, we will discuss factors that influence the expression of genes (such as epigenetic modification) and methods to study how gene expression is regulated and how genes interact.

The reader is assumed to have some basic knowledge of genetic terminology. Genetic information is encoded in DNA (deoxiribonucleic acid) molecules. The DNA code contains the units of genetic information we call genes. There is no real agreement on what exactly defines a gene, the definition has evolved along with the advances in science. Commonly used definitions of a gene are: 'a unit of inheritance' or 'a packet of genetic information that encodes a protein or RNA'. The estimated number of genes in the human genome is also a subject of debate. Not too long ago it was predicted that the human genome contained around 100,000-150,000 genes (e.g., Liang et al., 2000). However, more recent estimates have gone down to 20,000-25,000 (International Human Genome Sequencing Consortium, 2004).

In humans, the DNA molecules are organised in 2 x 23 chromosomes: 22 pairs of autosomes and one pair of sex chromosomes. The genetic sequence as a whole is called the genome, and a location in the genome that for instance contains a gene or a genetic marker is referred to as a locus. A quantitative trait locus (QTL) is a locus that harbours a gene influencing a quantitative trait, i.e. a trait that varies on a quantitative scale. Assessment of the trait may be on an interval or ordinal scale, in which case an underlying quantitative liability is assumed.

The nuclei of nearly all human cells contain two versions of each chromosome, and therefore of each gene. The two corresponding chromosomes...
INTRODUCTION

Individual differences in psychological or behavioral traits can be explained by a combination of genetic and environmental differences between individuals. When a trait is said to be highly genetic, this means that a large proportion of the variance in the trait is explained by genetic factors, i.e. the effect of one or many genes that each have their influence on the expression of the trait. Environmental factors can range from intrauterine environment to the influence of the family environment, school, friends and many other unidentified non-genetic factors. In this chapter we will provide an overview of methods used to model the contribution of genes and environment to variance in a trait or a set of traits, and to localize and identify the regions of the genome that may be involved. The area of research that focuses on quantifying genetic effects is called behavior genetics or genetic epidemiology. Genes can be localized and identified with genetic linkage and association methods. Finally, we will discuss factors that influence the expression of genes (such as epigenetic modification) and methods to study how gene expression is regulated and how genes interact.

The reader is assumed to have some basic knowledge of genetic terminology. Genetic information is encoded in DNA (deoxiribonucleic acid) molecules. The DNA code contains the units of genetic information we call genes. There is no real agreement on what exactly defines a gene, the definition has evolved along with the advances in science. Commonly used definitions of a gene are: ‘a unit of inheritance’ or ‘a packet of genetic information that encodes a protein or RNA’. The estimated number of genes in the human genome is also a subject of debate. Not too long ago it was predicted that the human genome contained around 100,000-150,000 genes (e.g., Liang et al., 2000). However, more recent estimates have gone down to 20,000-25,000 (International Human Genome Sequencing Consortium, 2004).

In humans, the DNA molecules are organised in 2 x 23 chromosomes: 22 pairs of autosomes and one pair of sex chromosomes. The genetic sequence as a whole is called the genome, and a location in the genome that for instance contains a gene or a genetic marker is referred to as a locus. A quantitative trait locus (QTL) is a locus that harbours a gene influencing a quantitative trait, i.e. a trait that varies on a quantitative scale. Assessment of the trait may be on an interval or ordinal scale, in which case an underlying quantitative liability is assumed.

The nuclei of nearly all human cells contain two versions of each chromosome, and therefore of each gene. The two corresponding chromo-
somes are called homologous chromosomes. One is received from the mother, the other from the father. In addition, a small amount of DNA is contained in the (maternally inherited) mitochondria. Although most of the human DNA sequence is identical in all individuals, at some loci different versions of the sequence occur. These variants are called alleles. The word allele can refer to a gene variant, but also to versions of a genetic marker or any other fragment of DNA sequence. Individuals who carry the same allele at both homologous chromosomes are called homozygous. Individuals with two different alleles are heterozygous. The two alleles together, either at one or at multiple loci, make up a person’s genotype. The term haplotype is used to indicate a combination of alleles at multiple loci that an individual receives from one parent (Ott, 1999). It usually refers to a combination of alleles transmitted close together on the same chromosome. Finally, the observed characteristics of an individual are called phenotypes.

Alleles affecting quantitative traits can exert their effect in various ways. When the alleles act independently, the effects simply add up, in which case we speak of additive genetic effects. When the effect of one allele depends on the effect of another, i.e. there is an interaction between them, they are referred to as non-additive effects. There are several forms of non-additivity. Interactions between two alleles at the same locus are referred to as dominance. When the interaction is between alleles at two different loci, it is referred to as epistasis. An excellent online tutorial by Shaun Purcell that addresses additivity and dominance can be found on http://pngu.mgh.harvard.edu/~purcell/bgim/-index2.html#sgene.

In this chapter we will provide an overview of genetic epidemiological methods and developments, in three sections. The first part will describe the estimation of heritability, as well as some more advanced modeling based on twin methodology. In part II, methodology used to localize and identify genes will be discussed. Finally, in part III, we will focus on the gene expression and epigenetic modification of the DNA.
Part I: Estimating Heritability

It is often observed that human traits run in families. This is not only the case for diseases or physical appearance but can also apply to personality and behavior. The mere fact that a trait is familial, however, does not tell us whether the trait is heritable, since familial resemblance can also be the result of the influence of a shared family environment.

One method to investigate genetic influences is by studying adopted children and their biological and adoptive parents. Similarities between adopted children and their biological parents reflect genetic influences, whereas similarities between adopted children and their adoptive parents reflect the effects of the family environment. However, there are some disadvantages to adoption studies: adoptions are relatively rare, and adoptive children and parents cannot be assumed to be representative of the general population.

The Classical Twin Model

For this reason many studies use data from twins and their families. In twin studies, the resemblance between monozygotic (MZ) and dizygotic (DZ) twins is compared to estimate the contribution of genes, shared environment and non-shared environment to the variance in a trait. This is based on the fact that MZ twins share 100% of their segregating genes, whereas DZ twins share on average 50%. [Note that this percentage refers to the portion of the genome in which variation occurs, since >99% of the genome is identical between humans; this part is therefore entirely shared in both MZ and DZ twins.] In contrast, both MZ and DZ twins share the home environment. This means that differences between MZ twins must be due to non-shared environmental influences, whereas the extent to which MZ twins are more similar than DZ twins reflects the influence of genetic factors. Using these principles, the variance in a trait can be decomposed as due to additive genetic factors (A), common or shared environment (C) and non-shared environment or measurement error (E). In the absence of dominance or epistasis, the percentage of variance in a trait that is explained by additive genetic factors equals the heritability of the trait, which can be estimated by taking twice the difference between the MZ and DZ twin correlation: \( h^2 = 2(r_{MZ} - r_{DZ}) \).

When \( r_{MZ} > 2r_{DZ} \), there is evidence for a contribution of non-additive genetic influences, also referred to as genetic dominance (D), which also includes effects of epistasis. In this case, the percentage of variance explained by
A and D together is referred to as the broad-sense heritability, A alone is called the narrow-sense heritability.

The contribution of A, C, D and E to the trait variance can be estimated based on biometrical genetic theory. Discussing the biometrical model in detail is beyond the scope of this chapter, but it is the basis of a few important principles twin models are based on. For a detailed introduction see Falconer and Mackay (1996).

As explained above, the total phenotypic variance of a trait (P) can be decomposed into components explained by A, C, D and E: \( V_P = V_A + V_D + V_C + V_E \). We here assume that there is no interaction or correlation between genetic and environmental factors (the covariance between A and D is zero by definition).

The covariance between MZ twins is expressed as: \( \text{cov(MZ)} = V_A + V_D + V_C \). Since \( V_E \) is by definition non-shared variance, it cannot contribute to covariance of family members. \( V_C \) is, by definition, shared, and the genetic variance is also entirely shared because MZ twins are genetically identical. The expectation for the DZ twin covariance is expressed as: \( \text{cov(DZ)} = \frac{1}{2} V_A + \frac{1}{4} V_D + V_C \). On average \( \frac{1}{2} \) of the additive genetic variance is shared between DZ twins (and between non-twin siblings). In order to share non-additive variance, two relatives have to share both alleles of a gene, an event that occurs with a probability of \( \frac{1}{4} \) in DZ twins (or full siblings). Figure 2.1 shows a graphical representation of the model that arises from these principles.

**Structural Equation Modelling**

To estimate the contribution of all genetic and environmental factors to a trait and assess their significance, models can be evaluated and compared using structural equation modelling (SEM). The parameters of a model (which include means, variances and covariances) can be estimated using an optimization approach such as maximum likelihood estimation. The relative goodness-of-fit of different models can be assessed by calculating minus twice the log-likelihood (-2LL) of the data given the model, and comparing these values between models. By dropping or equating parameters, the fit of different models can be compared with a likelihood ratio test. Genetic structural equation modelling usually involves a multiple group design in which data from e.g. MZ and DZ twins are analyzed simultaneously and parameters (a, c, d and e in Figure 2.1) are constrained to be equal across groups to ensure identification of the model. Usually, a fully saturated model that includes estimates for all parameters is tested first. Then the significance of parameters can be tested by constraining them to be zero. For instance, it can be tested whether the C factor...
has an effect on the variance of the trait, by fixing the c path coefficient at zero and then comparing the original model to the constrained model. When dropping or equating parameters does not result in a significant deterioration of the model fit, this indicates the more parsimonious model fits the data as well as the more complex model. The best model is the most parsimonious model that still provides a good explanation of the observed data. Significance is determined based on the difference in -2LL between two models, which is asymptotically distributed as $\chi^2$. The degrees of freedom of the test are equal to the difference in the number of parameters. For very large samples alternative fit indices have been proposed, such as the RMSEA (Browne & Cudeck, 1993), the Bayesian Information Criterion (BIC; Schwarz, 1978) and the Akaike Information Criterion (AIC; Akaike, 1987).

**Twin models and categorical data**

In the case of a continuous variable, the trait is assumed to be normally distributed [which is indeed expected for traits that are affected by many genes (Fischer, 1918)]. Clearly, non-continuous phenotypes (e.g. presence or absence of a disorder, or categories representing levels of severity of a phenotype) are not normally distributed, and cannot be analyzed the same way. However, they may reflect a categorization of an underlying normally distributed trait. In this situation, a liability threshold model (Falconer, 1965) is often used. A threshold model assumes that the categories of a variable reflect an imprecise measurement of an underlying normal distribution of liability with a mean of zero and a variance of one. One or more thresholds (expressed as Z-scores) divide this distribution into discrete classes (e.g. affected vs. unaffected for a disease phenotype, or no symptoms/mild/moderate/severe for a trait measured on a continuous scale, such as a neuroticism or depression score). The area under the curve between two thresholds represents the proportion of cases within a category (Figure 2.2). The resemblance of relatives (e.g. twins) is expressed as tetrachoric or polychoric correlations, which represent the correlation of relatives on the liability dimension.
Univariate twin model. The path diagram shows the A, C, D and E factors for a twin pair, and the correlations between each of the factors for MZ and DZ twins. Following the tracing rules of path analysis (Wright, 1934), the phenotypic variance explained by each component is calculated as the squared path coefficient: the genetic variance for an individual is calculated as $a^2$, the shared environmental variance equals $c^2$, etc. The total variance is derived by summing the variance explained by the individual components: $a^2 + c^2 + d^2 + e^2$. The covariance between twins is calculated by tracing the path from twin 1, through the double-headed arrow, to twin 2. For instance, the genetic covariance between MZ twins equals $a \cdot 1 \cdot a = a^2$, whereas for DZ twins it equals $a \cdot 0.5 \cdot a = 0.5a^2$. The total covariance is calculated by adding up all paths contributing to the covariance (i.e. all paths which connect the two twins), which is $a^2 + c^2 + d^2$ for an MZ pair and $0.5a^2 + c^2 + 0.25d^2$ for a DZ pair. Note that when only data from twins reared together are available, it is not possible to estimate C and D at the same time, because there is not enough information; an ACDE model is not identified. Therefore, the twin correlations are used to decide whether an ACE or an ADE model is more plausible.
Figure 2.1
Univariate twin model. The path diagram shows A, C, D and E factors for a twin pair, and the correlations between each of the factors for MZ and DZ twins. Following the tracing rules of path analysis (Wright, 1934), the phenotypic variance explained by each component is calculated as the squared path coefficient: the genetic variance for an individual is calculated as $a^2$, the shared environmental variance equals $c^2$, etc. The total variance is derived by summing the variance explained by the individual components: $a^2 + c^2 + d^2 + e^2$. The covariance between twins is calculated by tracing the path from twin 1, through the double-headed arrow, to twin 2. For instance, the genetic covariance between MZ twins equals $a_1 a = a^2$, whereas for DZ twins it equals $a_0.5 a = 0.5a^2$. The total covariance is calculated by adding up all paths contributing to the covariance (i.e. all paths which connect the two twins), which is $a^2 + c^2 + d^2$ for an MZ pair and $0.5a^2 + c^2 + 0.25d^2$ for a DZ pair. Note that when only data from twins reared together are available, it is not possible to estimate C and D at the same time, because there is not enough information; an ACDE model is not identified. Therefore, the twin correlations are used to decide whether an ACE or an ADE model is more plausible.

Extensios of the classical twin model
The classical twin design can be extended to also include data from siblings, parents and spouses. The genetic similarity between non-twin sibling pairs is the same as the resemblance between DZ twins, i.e. on average 50% of the segregating genes. Adding data from one or more non-twin siblings to the model (often referred to as an extended twin design) results in a substantial increase in power to detect genetic and shared environmental effects (e.g., Posthuma & Boomsma, 2000).

The similarity between parents and children is 50% for additive genetic effects but 0% for dominance; since dominance reflects an interaction between two alleles at the same locus, to share these effects two individuals have to share both alleles. However, parents by definition transmit only one allele to their children.

Data from parents and spouses can be used to account for the effects of parental influence (i.e. cultural transmission) and assortative mating (i.e. phenotypic correlations between spouses; Fulker, 1982). An example of this method can be found in e.g. Distel et al. (2009), who investigated whether

Figure 2.2
Threshold models. In both cases a normal distribution of liability underlies the observed phenotypes, which have been categorized into discrete classes. A: single threshold model; this represents a disease phenotype with affected and unaffected individuals. B: multiple threshold model; in this case an ordinal variable with categories corresponding to different levels of severity, in this case ranging from no symptoms, via mild and moderate, to a severe phenotype.
cultural transmission from parents to offspring had an effect on borderline personality features. They found that cultural transmission did not play a role; however, there was some evidence for assortative mating, although this explained only a small amount of the variance in the trait.

**Multivariate models**

A useful extension of the models described above is to analyze multiple traits simultaneously. Bivariate or multivariate models can be used to quantify the genetic and environmental overlap in correlated traits, and explore the etiology of the association (or comorbidity) between traits. For example, it is possible to test whether the same genes affect different correlated traits, or whether a similar environment is responsible for the correlation.

In addition to the MZ and DZ twin correlations, a multivariate model also includes the phenotypic correlation between traits (within a person), and the cross-twin cross-trait correlation (the correlation between trait 1 in twin 1 and trait 2 in twin 2). The function of the cross-twin cross-trait correlations is similar to that of the regular twin correlations in a univariate model: if the cross-twin cross-trait correlation is higher in MZ than in DZ twins, this indicates the two traits share a genetic component, in other words, there is a genetic correlation between them. Shared and non-shared environmental correlations are calculated similarly. Figure 2.3 shows an example of a bivariate ACE model. The cross-twin cross-trait correlations are modeled by adding the cross-paths $a_{21}$, $c_{21}$, and $e_{21}$. If the $a_{21}$ path is significant, this implies that a genetic correlation is present, and similarly, significance of $c_{21}$ and $e_{21}$ indicates shared and non-shared environmental correlations, respectively. For instance, following the tracing rules of path analysis (Wright, 1934), the genetic covariance between phenotype 1 in twin 1 and phenotype 2 in twin 2 in DZ twins is given by $a_{11} \cdot 0.5 \cdot a_{22}$.

An example of a bivariate twin analysis is described in Chapter 7. In this analysis, the relationship between migraine and depression was investigated, to test the hypothesis that the often-reported comorbidity of these disorders is due to a shared underlying genetic factor. The phenotypic correlation between the two traits was estimated at .28. Most of the shared variance (54%) was explained by genetic factors, the remaining variance was due to non-shared environment. There was a significant genetic correlation between the traits ($r = .30$). Thus, it can be concluded that migraine and depression are in part influenced by the same genetic and non-shared environmental factors, but that the proportion of variance explained by this relationship is modest.
Chapter 2

Cultural transmission from parents to offspring had an effect on borderline personality features. They found that cultural transmission did not play a role; however, there was some evidence for assortative mating, although this explained only a small amount of the variance in the trait.

Multivariate Models

A useful extension of the models described above is to analyze multiple traits simultaneously. Bivariate or multivariate models can be used to quantify the genetic and environmental overlap in correlated traits, and explore the etiology of the association (or comorbidity) between traits. For example, it is possible to test whether the same genes affect different correlated traits, or whether a similar environment is responsible for the correlation.

In addition to the MZ and DZ twin correlations, a multivariate model also includes the phenotypic correlation between traits (within a person), and the cross-twin cross-trait correlation (the correlation between trait 1 in twin 1 and trait 2 in twin 2). The function of the cross-twin cross-trait correlations is similar to that of the regular twin correlations in a univariate model: if the cross-twin cross-trait correlation is higher in MZ than in DZ twins, this indicates the two traits share a genetic component, in other words, there is a genetic correlation between them. Shared and non-shared environmental correlations are calculated similarly. Figure 2.3 shows an example of a bivariate ACE model.

The cross-twin cross-trait correlations are modeled by adding the cross-paths $a_{21}$, $c_{21}$ and $e_{21}$. If the $a_{21}$ path is significant, this implies that a genetic correlation is present, and similarly, significance of $c_{21}$ and $e_{21}$ indicates shared and non-shared environmental correlations, respectively. For instance, following the tracing rules of path analysis (Wright, 1934), the genetic covariance between phenotype 1 in twin 1 and phenotype 2 in twin 2 in DZ twins is given by $a_{11} \cdot 0.5 \cdot a_{21}$.

An example of a bivariate twin analysis is described in Chapter 7. In this analysis, the relationship between migraine and depression was investigated, to test the hypothesis that the often-reported comorbidity of these disorders is due to a shared underlying genetic factor. The phenotypic correlation between the two traits was estimated at .28. Most of the shared variance (54%) was explained by genetic factors, the remaining variance was due to non-shared environment. There was a significant genetic correlation between the traits ($r = .30$). Thus, it can be concluded that migraine and depression are in part influenced by the same genetic and non-shared environmental factors, but that the proportion of variance explained by this relationship is modest.

Another application is the extension to longitudinal models. By measuring correlations between repeated measures at different time points, it can be determined whether stability over time is due to genetic or environmental factors. An example can be found in Bartels et al. (2004), who investigated the contribution of genes and environment to stability in internalizing and externalizing problem behavior in children aged 3 to 12 years old. They found that genetic factors were responsible for both stability and change over time, while a common set of shared environmental factors mostly accounted for stability in problem behavior across different ages. Non-shared environment played only a modest role in explaining stability or change in problem behavior.

Figure 2.3

Example of a bivariate twin model: a bivariate ACE model, with two twins and two phenotypes (P1 and P2).
Sex by genotype and age by genotype interaction

The influence of genetic and environmental factors may differ for males and females. Therefore it may be useful to first test a full model in which all parameter estimates are different for the two sexes, and then test whether the estimates for males and females can be constrained to be equal. In many situations means or thresholds have to be modeled separately for males and females, for instance because a trait (e.g. migraine or depression) is more prevalent in women. Apart from that, several different hypotheses can be tested:

1. The variance components (i.e. $V_A$, $V_D$, $V_C$ and $V_E$) are the same for males and females.
2. The variance components are proportionally the same in males and females, but in one sex the total trait variance is larger. A way to model this is by constraining all variance components in one sex to be a scalar multiple of the variance components in the other sex.
3. The variance components differ, for instance when a trait is more heritable in one sex than in the other. In this case, the variance components have to be estimated separately for men and women. Note that a decrease in heritability may arise for different reasons: the genetic variance can be the same in the two sexes, but the environmental variance could be larger in men than in women. Since heritability is expressed as a ratio (genetic variance over total variance) this would lead to a lower heritability estimate in men.

To test which of these models fits the data best, one starts with a full model in which all parameters are estimated separately for males and females. Then, by constraining the parameters step by step, it is tested whether parameter estimates differ significantly between men and women. This same method can be applied when data are available for different age groups (e.g. adolescents vs. adults), to test whether estimates of heritability differ depending on age. The actual implementation of the model depends on whether data have been collected in a cross-sectional design in subjects of different ages, or in a longitudinal study in which the same subjects are measured repeatedly across time.

Finally, when data from DZ opposite sex (DOS) twins are available, it is possible, in addition to quantitative differences, to also test whether qualitative sex differences are present (i.e. whether different genes affect the trait in males and females). This is tested by estimating the correlation between the latent
genetic factors in DOS twins, while this correlation remains fixed at 0.5 in the same sex pairs. If the correlation in DOS twins is significantly lower than 0.5, this is an indication that the genetic factors affecting males and females are (partly) different. It is also possible that different environmental factors influence a trait in men and women. In this case the correlation between the C factors (see Figure 2.1) would be estimated in DOS twin pairs (or in opposite-sex siblings). It is not possible to estimate the correlations for genetic and shared environmental factors simultaneously using a classical twin design, as there is only one data point available that is informative for this test.

**Genotype by environment interaction**

The expression of genes may also depend on environmental factors - sometimes referred to as moderators. For example, the expression of the genotype may be more clearly seen in a permissive environment. An interesting case of gene-environment interaction (GxE) was found in a study by Boomsma et al. (1999), who observed that a religious upbringing reduced the influence of genetic factors on disinhibition, one of the dimensions of the Sensation Seeking Scales. In a study of female twins Heath et al. (1998) found that being in a marriage-like relationship served as a protective factor by reducing the impact of a genetic liability to depression.

When it is known which genes influence a particular phenotype it is also possible to test for interaction of the environment with a specific gene variant. In a famous study, Caspi et al. (2003) investigated the association between the serotonin transporter gene and depression in individuals who had experienced stressful life events and individuals who had not. It was found that stressful life events were associated with depression, but only in individuals who carried at least one copy of the short allele of the serotonin transporter gene. The strongest effect was observed in individuals who carried two copies of the short allele, while the effect was non-significant in carriers of two long alleles. As spectacular as these results were, these days it is thought they may have been chance findings, since few studies since have succeeded in reproducing them. A large meta-analysis of the many replication studies failed to show significant evidence of either a main effect of the serotonin transporter gene or an interaction between this gene and stressful life events (Risch et al., 2009).
PART 2: GENE-FINDING

Once it has been established that a trait is heritable, the next step is to find the genes involved. The two primary statistical methods for gene-finding are linkage and association. Unlike the methods described above, linkage and association require the collection of DNA samples and the measurement of genotypes.

Linkage analysis is a method that localizes regions possibly influencing the trait of interest by using pedigree information. In short, the objective is to determine whether relatives who are phenotypically similar, are also genotypically similar in a particular region of the genome. If this is the case, this region may harbour a gene involved in the trait of interest. Linkage is based on the principle that two loci that are physically close together (e.g. an observed fragment of DNA and an unobserved disease locus) are more likely to be co-inherited. How this works will be discussed in more detail below. Because the information in a linkage study comes from the pedigree structure, it is necessary to collect family data.

Association analysis can go one step further: not only can the location of the involved regions be determined but also which genetic variant (allele) is associated with the phenotype. In other words: do individuals with a certain phenotype have a different frequency of allele X than individuals who do not have this phenotype? This can be tested with a straightforward chi-squared or regression test. Association studies have a higher resolution than linkage studies and have often been used to follow up promising linkage results. As we will see, using family data has certain advantages; however, association studies can also be performed using data from unrelated individuals.

MARKERS

Because - due to technical and financial limitations - it is currently not feasible to characterize the entire human DNA sequence in large numbers of individuals, gene-finding studies rely on markers. Markers are genetic variants (also called polymorphisms) with a known location which can be used as indicators of the approximate location of the real, usually unmeasured locus of interest. When we say an individual is genotyped for a linkage or association study, this means their DNA is characterized at a selected number of marker loci, either in a specific region (in candidate gene studies), or throughout the genome (in genome-wide studies).
Several types of markers are used in gene-finding studies. Single nucleotide polymorphisms (SNPs) are single base pairs with two variants (e.g. some individuals have an A, others have a C). Theoretically (if single base pair mutations have occurred multiple times at the same locus) there can be 4 variants (A, C, T and G), but for practical reasons only SNPs with two variants are selected for gene-finding studies. Microsatellites are sequence length polymorphisms that consist of a varying number of repeats of a short (usually 1-4 bp) sequence of DNA, e.g. ‘CACACACACACACA’. A third and more recently recognised type of polymorphism is the copy number variant (CNV). CNVs are DNA fragments ranging from kilobases (Kb) to even megabases (Mb) in size, of which different numbers of copies are present in different individuals.

**Parametric linkage**

Broadly speaking, two types of linkage analysis can be distinguished: parametric and nonparametric linkage. Parametric (or model-based) linkage requires the specification of a genetic model, i.e. allele frequencies and penetrances (3 parameters specifying the probability that an individual expresses the phenotype given 0, 1 or 2 copies of the risk allele) Genotype and phenotype data from multiple generations are required to perform this type of analysis.

An important concept in parametric linkage analysis is the recombination fraction. Recombination occurs when during meiosis the maternal and paternal chromosome cross over, break and rejoin, resulting in gametes with chromosomes that are a combination of the maternal and paternal chromosome.

The recombination fraction, used in linkage analysis, is the probability that the alleles at two loci are recombinant (i.e. an odd number of recombination events has occurred between them). This depends on the distance between the loci. When two loci are located on different chromosomes, or on the same chromosome but far apart, the probability of the individual being a recombinant is around 50%, i.e. the recombination fraction (θ) is 0.5. The smaller the distance between the two loci, the lower the probability of a recombination event between them, and the lower θ will be, with θ = 0 indicating perfect linkage.

To test for linkage, a genetic model is assumed, and the likelihood of the observed pedigree data under the alternative hypothesis of linkage (θ < 0.5) is compared to the likelihood under the null hypothesis of no linkage (θ = 0.5) between the measured marker locus and the hypothetical trait locus. The result of this test is expressed as the logarithm of odds, called the LOD score. The
higher the LOD score, the stronger the evidence for linkage. A detailed discussion of parametric linkage methods can be found in Ott (1999).

Parametric linkage is most suited for traits that are influenced by a single gene and follow a relatively simple pattern of inheritance, because in this situation it is relatively easy to specify a genetic model. A good example of a successful parametric linkage study is described by Joutel et al. (1993), who used data from two large multigenerational families to map the first locus for familial hemiplegic migraine (the FHM1 locus) to chromosome 19. A few years later, Ophoff et al. (1996) identified several mutations in a gene in this area (CACNA1A) which caused the FHM phenotype. However, for many behavioral and psychological traits, specifying the correct genetic model is not straightforward.

**Nonparametric (model-free) linkage**

Most behavioral and psychological phenotypes are complex, i.e. they are influenced by many genes that each have a small effect. In this case it is difficult to specify a genetic model. Therefore, complex traits are usually analysed using non-parametric (also called model-free) linkage techniques. The non-parametric approach does not require the specification of a genetic model. In short, in non-parametric linkage, it is tested whether relatives with similar phenotypes also have similar genotypes. Genotypic similarity is expressed in a measure called identity by descent (IBD). Two alleles are said to be IBD if they not only have the same DNA sequence (referred to as identity by state, or IBS), but were also inherited from the same ancestor. Because there are two alleles for each locus, a pair of individuals can share 0, 1 or 2 alleles IBD. The expected probabilities for these values are $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$, respectively (Figure 2.4).

To test for linkage, the IBD values for all pairs of related individuals in the sample are estimated. For IBD estimation, the availability of parental genotypes greatly increases the accuracy of the estimates. For this reason, parental genotype data are used in linkage analysis, even when the actual LOD scores are based on data from siblings only.
Several algorithms have been developed for the estimation of IBD values. The Elston-Stewart algorithm (Elston & Stewart, 1971) is suited for analysis of very large pedigrees but only for a limited number of markers at a time, because the complexity of the calculations increases exponentially with the number of markers. The Lander-Green algorithm (Lander & Green, 1987) is better suited to handle the large numbers of markers included in most modern linkage studies, but is limited to smaller pedigrees. A useful discussion of IBD estimation can be found in, e.g., Ferreira (2004).

**Haseman-Elston regression**
One of the first non-parametric linkage methods based on IBD estimation was introduced by Haseman & Elston (1972). This method is now known as Haseman-Elston (HE) regression. The idea was to take the squared difference in trait values for each sibling pair and regress it on the estimated IBD values at a given marker locus. There is evidence for linkage when high IBD values are
Chapter 2

associated with strong phenotypic similarity (i.e. small squared trait differences). Thus, a significant negative regression slope indicates the presence of linkage. A drawback of HE regression is that fairly large samples are needed for sufficient statistical power to detect linkage. One method to increase power is by selecting only the most extreme cases from a population (Carey & Williamson, 1991; Dolan & Boomsma, 1998). This is possible because HE regression has the advantage that it does not rely on assumptions about the trait distribution.

Several extensions to HE regression have been proposed through the years, which improve power by using not only the squared trait differences but also the squared trait sum (e.g., Sham et al., 2002).

**Variance components linkage**

A non-parametric linkage method developed in the 1990’s is based on variance components (VC; e.g., Almasy & Blangero, 1998; Amos, 1994). VC linkage is based on an approach similar to that described in the section about heritability estimation. In addition to the genetic and environmental components A, C, D and E, we can model the effect of a specific QTL (Q), using IBD estimations. Figure 2.5 shows a model that incorporates A (background genetic effects), Q (QTL effect) and E (environment). The correlation between the QTL factors of DZ twins and siblings equals the estimated proportion of alleles IBD, which is referred to as \( \hat{\pi} \) (‘pi-hat’).

To test whether there is significant linkage at a certain locus the path coefficients for the Q-factor (q) are constrained to be zero. A significant deterioration of the model fit is taken as evidence for linkage. This procedure is repeated for all loci and significance levels should be adjusted accordingly. The advantage of VC linkage is that, unlike HE regression, it can be used with any type of pedigree, and it is generally more powerful. An important disadvantage, however, is its reliance on the assumption of normality of the trait distribution. Hence, the analysis of data from selected samples with variance components linkage is more involved than when HE regression is used.
Chapter 2

46

associated with strong phenotypic similarity (i.e. small squared trait differences). Thus, a significant negative regression slope indicates the presence of linkage. A drawback of HE regression is that fairly large samples are needed for sufficient statistical power to detect linkage. One method to increase power is by selecting only the most extreme cases from a population (Carey & Williamson, 1991; Dolan & Boomsma, 1998). This is possible because HE regression has the advantage that it does not rely on assumptions about the trait distribution.

Several extensions to HE regression have been proposed through the years, which improve power by using not only the squared trait differences but also the squared trait sum (e.g., Sham et al., 2002). Variance components linkage A non-parametric linkage method developed in the 1990’s is based on variance components (VC; e.g., Almasy & Blangero, 1998; Amos, 1994). VC linkage is based on an approach similar to that described in the section about heritability estimation. In addition to the genetic and environmental components A, C, D, and E, we can model the effect of a specific QTL (Q), using IBD estimations.

Figure 2.5 shows a model that incorporates A (background genetic effects), Q (QTL effect) and E (environment). The correlation between the QTL factors of DZ twins and siblings equals \( \hat{\pi} = \text{IBD}/2 \), whereas the correlation between the background genetic factors of siblings or DZ twins is .5. If the mode of inheritance of the trait is largely unknown, the remaining familial variance that cannot be attributed to Q can also be modelled as simply ‘familial’. If data on MZ twins are also available the familial variance can be decomposed into A and C. Note that MZ twins do not contribute any information to detect linkage (as they are perfectly correlated for all QTLs).

The affected sib pair method

For disease phenotypes (i.e. affected vs. unaffected) a commonly used linkage method is the affected sib-pair (ASP) test. In an ASP design, it is tested whether sibling pairs who are both affected for a disorder share more alleles IBD than expected in the absence of linkage (in which case the distribution should be roughly \( \frac{1}{4}, \frac{1}{2}, \frac{1}{4} \) for IBD values of 0, 1 and 2, respectively).

As mentioned earlier, parametric linkage was successfully used to map a gene for familial hemiplegic migraine, which is a monogenic form of migraine. Common migraine, however (i.e. MO and MA) is polygenic in nature. Therefore, nonparametric methods are more suited to investigate this disorder. Chapter 8
of this thesis describes a study in which an affected sib-pair approach was used to analyse migraine in a sample of Dutch twins and their parents and singleton siblings. Suggestive linkage was detected on chromosomes 1, 13 and 20, and a previous finding by Nyholt et al. (2005) on chromosome 5 was replicated.

**The multiple testing problem**

These days linkage is usually performed genome-wide, in an exploratory fashion. Because in a genome-wide linkage study several hundreds of markers are tested simultaneously, a multiple testing burden is inevitable. Therefore, stringent significance thresholds have to be applied. Based on a simulation study, Lander & Kruglyak (1995) proposed using a LOD score of 3.6 to indicate significant linkage, which corresponds to a p-value of $2 \times 10^{-5}$ and should be roughly equivalent to a genome-wide significance level of 5%. This has become a widely used threshold to define significance in linkage studies. Alternatively, permutation or simulation approaches using the observed data can be used to determine empirical p-values. This has the advantage that no assumptions need to be made about the null distribution of the linkage statistic.

**Association**

In an association study, it is tested whether a particular allele or genotype is more prevalent in individuals with a certain phenotype. For instance, do individuals with allele C at a given SNP have a higher depression score than individuals with allele A?

Association analysis can be performed in unrelated individuals or in family-based samples. Studies in unrelated samples are often set up as case-control studies: allele or genotype frequencies are compared between a selection of cases and a group of matched controls. It is also possible to test for association with a continuous phenotype: in this case mean trait values are compared between individuals with different genotypes. The advantage of case-control association studies is the relative ease of collecting samples and the straightforward statistical tests that can be used. The disadvantage, however, is that the presence of an underlying population substructure can lead to spurious results, a phenomenon referred to as ‘population stratification’. This phenomenon is illustrated in a famous paper by Hamer & Sirota (2000). The paper describes a hypothetical study in a student population consisting of Caucasian and Asian subjects, in which a gene is identified for eating with chopsticks. However, this is not a true association, but the result of the fact that, for all sorts of reasons, allele frequencies can differ between the two
populations. The two populations happen to also differ in terms of eating with chopsticks, which is entirely culturally determined. However, from the association analysis it falsely appears that the gene has something to do with the chopsticks.

One way to deal with stratification issues is by using a family-based association test, such as the haplotype relative risk (Falk & Rubinstein, 1987; Terwilliger & Ott, 1992), or the transmission disequilibrium test (Spielman et al., 1993), which use data from heterozygous parents and affected children to determine which parental alleles are transmitted to an affected child and which are not. Thus, the non-transmitted alleles serve as ‘internal’ control genotypes, which eliminates the need for external controls and the risk of stratification issues. The disadvantage is that family-based samples are more difficult to collect, because both parents have to be present (which can be particularly challenging for late-onset phenotypes such as Alzheimer’s disease or ageing).

An alternative approach, suitable for quantitative traits, was developed by Fulker et al. (1999). With this method, which uses data from sibling pairs, the effects of genes on phenotypic means are partitioned into a between and within-family component. A within-family association test is not affected by population stratification because siblings within a family belong to the same stratum. Thus, it is tested whether an allele is associated with the phenotype in siblings within the same family, whether they are associated in siblings from different families, and whether the effect size of these tests is the same. If the gene effect is different between families than within families, there is evidence for population stratification. If, however, the within-family effect alone is significant, regardless of the between-family effect, this means there is still evidence for a true association effect, not due to population substructure (Fulker et al., 1999). This method has been implemented in the QTDT program (Abecasis et al., 2000).

In situations where unrelated individuals are used for association analysis, other methods are available to assess and control for population stratification, such as calculating the genomic inflation factor and applying genomic control. These will be discussed in more detail later.

**Linkage Disequilibrium**

An important concept in association studies is the phenomenon of linkage disequilibrium (LD). When two loci are in linkage equilibrium, the genotype at locus 1 is independent of the genotype at locus 2. This is usually the case when loci are on different chromosomes or far apart on the same chromosome.
However, if two loci are close together and over generations little recombination has taken place between them, the genotype at locus 1 may be associated with the genotype at locus 2. Therefore, when an association is found, this can be due to either direct or indirect association. In the case of direct association, the association signal comes from the actual causal variant. An association that arises because the marker is in LD with the causal variant, it is called indirect association.

**CANDIDATE GENE STUDIES VS. GENOME-WIDE ASSOCIATION**

Until recently, association studies focused on smaller candidate regions. Based on existing knowledge (e.g. theories about biochemical pathways or evidence from linkage studies), candidate genes were identified and genotyping was restricted to the region of interest. Good examples are association studies of the serotonin receptor and transporter genes in both depression and migraine studies. Both conditions are often successfully treated with drugs that interact with the serotonergic system (selective serotonin reuptake inhibitors [SSRIs] and triptans, respectively), suggesting a possible causal involvement of serotonin in the etiology of the disorders. However, in spite of the large number of studies conducted, it has proven difficult to unequivocally demonstrate a role of serotonin receptor or transporter genes in the pathogenesis of depression (Anguelova et al., 2003; Risch et al., 2009). A similar conclusion can be drawn for migraine (Colson et al., 2007). Although there is limited evidence for a possible role of certain serotonin-related genes in migraine and depression, the majority of candidate-gene association studies have returned negative results. This may be illustrative of the main weakness of the candidate gene approach: usually our knowledge about the pathways involved is very limited, making it very difficult to determine which genes are good candidates.

Due to the availability of faster and cheaper genotyping techniques it has now become feasible to genotype enough markers (from 300.000 up to 1 million) to cover most of the common variation in the entire human genome, and perform genome-wide association analysis (GWA). Several companies (e.g. Illumina, Affymetrix) produce pre-designed SNP chips that include a selection of carefully chosen ‘tag SNPs’. Tag SNPs are selected based on LD patterns, in such a way that a minimum number of SNPs captures a maximum amount of genetic variation in the population it is designed for. In contrast with candidate gene studies, a GWA study is exploratory in nature; no prior hypothesis about the location of causative genes is necessary. Indeed, many of the associations identified through GWA studies to date were not previously regarded as...
candidates, which demonstrates the use of exploratory gene-finding studies (Manolio & Collins, 2009).

**The multiple testing problem (2)**

Due to the large numbers of markers used, the multiple testing burden in a GWA study is even larger than in a linkage study, which makes it crucially important to use appropriate significance thresholds. The exact multiple testing burden depends on the set of SNPs included in the study and on the population studied. For instance, African populations are known to have less LD and more SNPs, and therefore the multiple testing burden will be higher than in a European population. Several authors have proposed cut-off values for significance in GWA studies. Pe’er et al. (2008) recommended multiplying the nominal p-value by the genome-wide testing burden, which, according to their calculations is roughly half a million tests when all common SNPs are tested in a European (Hapmap CEU) population. To obtain a genome-wide significance level of 5%, this means a nominal threshold of \( P = 1 \times 10^{-7} \) should be used. Dudbridge & Gusnanto (2008) used a permutation approach to estimate the genome-wide significance threshold in the UK Caucasian population. They estimated that genome-wide significance at the 5% level corresponded to a nominal P-value of \( 7.2 \times 10^{-8} \), and state that any P-value below \( 5 \times 10^{-8} \) can be considered “convincingly significant”.

It should be noted that even the use of strict significance thresholds has not been able to avoid that many candidate-gene association studies have produced results that could not be replicated, possibly because many of them were false positive findings (Hirschhorn et al., 2002). Since the credibility of a finding increases considerably when it is replicated in multiple independent samples, it is now a common requirement for GWA studies that results be replicated internally (i.e. in an independent sample described in the same study), in order to be published.
Chapter 2

**Figure 2.6**

Examples of Q-Q plots. The expected distribution of p-values (x-axis) is plotted against the observed distribution (y-axis). For convenience, p-values in a GWA study are often shown on a logarithmic scale, i.e. \(-\log_{10}(P)\). **A:** The observations closely follow the expected distribution (shown in grey), indicating there is probably no association and no inflation of the distribution either. **B:** Inflation across the whole distribution, which may indicate population stratification. **C:** An excess of small p-values in the tail of the distribution, possibly indicating some true associations.

**Genomic inflation**

As mentioned before, population substructure is a factor that can lead to spurious results in an association study. Therefore, in the design of a GWA study it is important to carefully select the individuals to be genotyped to avoid problems related to stratification within the sample. Once the data have been collected, it is common practice to run some quality control checks to scan for potential problems. A good way to get a first impression of the results is by creating a quantile-quantile plot (also called Q-Q plot). In a Q-Q plot, the expected distribution of p-values is plotted against the observed distribution (Figure 2.6). Under the assumption of no true association signals, this should result in a straight diagonal line. An excess of small p-values, resulting in a deviation in the tail of the distribution, may indicate true association signals. However, if the observed findings are inflated (i.e. show an excess of small p-values) across the entire distribution, this may indicate population stratification (McCarthy et al., 2008). The extent to which the distribution is inflated can be expressed in a statistic called the genomic inflation factor, \(\lambda\) (lambda), which is calculated as the median \(\chi^2\) of the observed distribution, divided by the median
\( \chi^2 \) of the expected distribution. Ideally, \( \lambda \) should approach a value of 1. Based on the value of \( \lambda \), the test statistic can be rescaled to correct it for inflation. This procedure is called genomic control (Devlin & Roeder, 1999).

**Meta-analysis**

GWA studies have shown to be effective and associations have been successfully identified for quite a number of human traits, such as Crohn’s disease (Barrett et al., 2008), type 2 diabetes (Zeggini et al., 2008), bipolar disorder (Ferreira et al., 2008) and obesity (Lindgren et al., 2009). However, it has become clear that for most complex traits the observed effects are small and therefore very large samples are needed. Visscher (2008) estimated that to detect a variant that explains 0.1-0.5\% of the variance in a quantitative trait (which may be a realistic effect size for genes affecting complex traits), tens of thousands of individuals are necessary for sufficient power. Since no single study has the budget to collect these enormous amounts of data, it is a necessity to combine GWA studies. For this reason, large consortia have been formed in recent years to enable meta-analyses of GWA results (e.g., Barrett et al., 2008; Lindgren et al., 2009; Zeggini et al., 2008). In a meta-analysis, the results of multiple individual studies are combined into one overall test statistic. Two types of meta-analysis can be distinguished: methods that assume fixed effects and those that assume random effects. Fixed effects methods assume there is one common effect in all studies (homogeneity), and that between-study variability is due to chance. Two frequently used fixed-effects methods are the inverse-variance weighted method and the pooled \( Z \)-score method. The inverse-variance weighted method pools the betas and standard errors from all studies, weighting each study by the inverse of the variance of beta. The outcome is an effect estimate for each SNP, pooled across all studies. This method is most suitable when the phenotype is measured on the same scale in all studies, so that beta can be interpreted the same way for all samples. The pooled \( Z \)-score method does not pool effect sizes but \( Z \)-scores, weighted by sample size. It provides information on the direction and significance of the pooled effect, but not about the effect size. This method is more appropriate when the phenotype is not measured on the same scale across studies and hence the effect sizes are not directly comparable.

In cases where different genetic effects are expected across studies (heterogeneity), for instance because the populations have a different genetic background, random effects methods are more appropriate. Various metrics are available to assess the presence of heterogeneity, such as Cochran’s Q statistic or I\(^2\) (Kavoura & Ioannidis, 2008). The main drawback of random effects
methods, however, is that they are more conservative and thus have low power compared to fixed effects models. A useful practical guideline for meta-analysis of genome-wide association studies is provided by de Bakker et al. (2008).

**HapMap & Imputation**

One problem in meta-analysis of GWA results is the fact that different studies use different SNP chips, which tend to be largely non-overlapping. As a consequence, the number of SNPs genotyped in all studies is limited. This can be overcome by imputing the genotypes of SNPs that were not measured, using data generated by the International HapMap project (2003; www.hapmap.org). The HapMap project was launched in 2002 with the purpose of creating a ‘haplotype map’ of the human genome that describes common patterns of genetic sequence variation. In phase I and II of the project 270 individuals from 4 populations (European, Nigerian, Japanese and Han Chinese) were genotyped to obtain information on more than 3.1 million SNPs. In phase III the project was expanded to include data from another 7 populations.

The HapMap data can be used to infer a missing genotype at one marker from available genotypes at other markers. This is possible because, due to the presence of LD, only a limited number of haplotypes frequently occur in the population, even though theoretically much more variation would be possible. To infer missing genotypes, a genotyped individual is compared to a HapMap reference sample. Because the LD structure in the HapMap sample is known it can be determined, given the available genotypes, what the most likely genotype is for the missing SNP. For instance, if all reference individuals with a certain haplotype have a C allele at SNP $X$, and SNP $X$ is in high LD with this haplotype, an individual with the same haplotype but a missing genotype at SNP $X$ is highly likely to also have a C allele.

Clearly there is some uncertainty involved in determining the most likely genotype for a missing SNP. For this reason, imputation programs calculate a probability for each possible genotype, and provide a quality measure that indicates how reliable the imputation is for each SNP, so that in the analysis stage, the researcher can decide to remove SNPs that were poorly imputed. In addition, the probability scores for the different genotypes can be used to account for the uncertainty of the imputations.

One limitation of the HapMap is that it covers only common variation. Therefore, if a trait is primarily influenced by rare alleles, associations will not be detected using the HapMap SNPs. The aim of the more recently started 1000 Genomes Project (www.1000genomes.org) is to provide coverage of the rarer
variants as well, and to provide a more detailed map of the human genome. In order to do this, whole genomes of approximately 1200 individuals will be sequenced (i.e. their entire DNA sequence will be determined).

**BOX: Genetics software on the internet**

**VARIOUS TYPES OF ANALYSIS:**
Merlin (Abecasis et al., 2002): http://www.sph.umich.edu/csg/abecasis/merlin/
For various types of parametric and non-parametric linkage, and association analysis in family data.

Mx (Neale et al., 2003): http://www.vcu.edu/mx/
Package for structural equation modelling, especially suitable for twin modelling and variance components linkage analysis.

**GENOME-WIDE ASSOCIATION:**
Plink (Purcell et al., 2007): http://pngu.mgh.harvard.edu/~purcell/plink/
GenABEL (R package) - http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/

**IMPUTATION:**
MACH (Li & Abecasis, 2006): http://www.sph.umich.edu/csg/abecasis/mach/
IMPUTE (Marchini et al., 2007): http://mathgen.stats.ox.ac.uk/impute/impute.html

**ANALYSIS OF IMPUTED DATA:**
SNPTEST (Marchini et al., 2007):
http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html
ProbABEL: http://mga.bionet.nsc.ru/~yurii/ABEL/

**META-ANALYSIS:**
METAL: http://www.sph.umich.edu/csg/abecasis/metal/
MetABEL (R package): http://mga.bionet.nsc.ru/~yurii/ABEL/

**OTHER USEFUL WEBSITES:**
Shaun Purcell’s behavioral genetic interactive modules:
http://pngu.mgh.harvard.edu/~purcell/bgim/
Greg Carey’s interactive learning exercises on behavior genetics:
http://psych.colorado.edu/~carey/hgss/hgssapplets/hgssapplets.htm
Mx script library (example scripts for twin modelling):
http://www.psy.vu.nl/mxbib/
Chapter 2

**Part III: Beyond Gene-finding**

A person’s phenotype depends on more than simply the genetic code. Genes exert their effects through their products, usually proteins. For proteins to be produced, a gene has to be expressed. The main steps in gene expression are transcription and translation. During transcription, the DNA molecule serves as a template to construct an RNA copy of itself (an RNA molecule resembles DNA but contains Uracil (U) instead of Thymine (T) bases, and is single-stranded). The RNA codes for a sequence of amino acids, together forming a protein. The construction of a protein, based on the RNA code, is called translation.

The expression of genes is affected by various factors, such as epigenetic modifications (see below) and regulation by other genes or transcription factors (proteins that bind to DNA, thereby controlling the expression of genes). In this last section, we will discuss the effects of epigenetic modification on gene expression, and the use of genome-wide expression data in gene-finding studies. Finally, a closely related area of research is the study of interactions between genes in biological networks and pathways. Identifying these pathways is an important step from statistical linkage or associations to understanding the biology underlying human traits and diseases.

**Epigenetics**

Epigenetics is the study of heritable changes in gene expression that are unrelated to changes in the DNA sequence. Epigenetic changes are caused by chemical modifications that affect the expression of genes. There are two types of modification that cause epigenetic changes: DNA methylation and histone modification. DNA methylation is the addition of a methyl group to a cytosine base that is followed by guanine (a so-called CpG site, where the p refers to the phosphodiester bond that connects two bases). CpG sites tend to occur in large repetitive sequences which are highly methylated, or in short CpG-rich DNA stretches called CpG islands, which are mostly unmethylated. CpG islands frequently overlap with the promoter region of genes (i.e. a region close to the gene where the transcription process is initiated). It is thought that methylation affects gene expression by controlling whether or not proteins that affect transcription can bind to the DNA (Jaenisch & Bird, 2003).

The second type of alteration is the modification of histones. Histones are the proteins around which DNA molecules are wrapped. There are various types of chemical modification of histones, including methylation and modifications affecting how densely the DNA is ‘packed’. A tightly packed structure of the
DNA prevents gene expression, whereas in relaxed DNA gene expression is active.

One might say that the epigenome has a lifecycle. After fertilization, most of the DNA is demethylized and a new wave of methylation occurs. This methylation pattern is inherited from parent to daughter cells during cell division, providing what might be called an ‘epigenetic memory’. Later in development, tissue-specific changes in methylation occur, which aid the differentiation of different cell types. At present, not much is known about how these changes occur (Feinberg, 2008). An interesting aspect of this phenomenon is that epigenetic changes are easier to reverse than genetic mutations, which may offer possibilities for the treatment of disease with drugs (e.g., Smith et al., 2007).

An additional factor that influences methylation patterns during the lifespan is the environment. Diet, for instance, has been suggested as an environmental factor that influences epigenetic processes. Diet-mediated epigenetic effects have been implicated in a variety of conditions, such as cancer, cardiovascular disease, but also depression and other psychiatric disorders (Van den Veyver, 2002). In recent years, it has become clear that epigenetics may explain part of the differences observed in genetically identical MZ twins. These differences will be part of the non-shared environmental component in a twin study. An interesting study in MZ twins showed that twins who were older, had more different lifestyles and spent less of their lifetimes together displayed more different epigenetic profiles than younger twins who shared most of their environment and lifestyle (Fraga et al., 2005). On the other hand, Heijmans et al. (2007), who combined an epigenetic study with a classical twin design, found that most of the variation across individuals in DNA methylation at the locus they investigated (IGF2/H19) could be attributed to heritable factors. The influence of environmental factors did not increase with age, suggesting at least some loci are relatively unaffected by age-related changes in methylation.

**Gene expression**

The genome-wide study of gene expression is a rapidly developing area of research. To measure gene expression, the transcript (RNA) content of a tissue sample is analyzed to determine which genes are being transcribed and in which quantities. One application of gene expression analysis is to combine it with the regular GWA approach. In this type of study, gene transcript abundance is treated as a phenotype, and can be mapped to genomic loci, called ‘expression
QTLs’ (eQTLs). This approach identifies markers that are associated with the expression of a gene and is useful to identify genetic variants that regulate the expression of other genes (Gilad et al., 2008). An example of how this might work is the situation where a strong association signal is found with an area that contains no genes (a so-called gene desert), a phenomenon that is regularly observed (Manolio & Collins, 2009). This region may harbour some regulatory sequence that influences the expression of a gene located at some distance from the associated SNP. An expression study might reveal this mechanism by detecting an association between the SNP in the gene desert and the expression level of the distant gene, which would otherwise go unnoticed.

A complicating factor in the collection of expression data is that expression levels differ depending on the type of tissue. Ideally, gene expression is measured in the tissues involved in the disease or trait of interest, however, in many cases (e.g. brain disease) it is not an easy task to obtain the right tissue samples in sufficient quantities. One possible solution could be to use more easily accessible tissues as a surrogate for the tissue of interest. For instance, Sullivan et al. (2006) compared gene expression in whole blood and 16 different tissues from the central nervous system (CNS) to assess the feasibility of using whole blood samples as a surrogate for brain tissue samples. They concluded that, although imperfect, there is a correlation between CNS and whole blood gene expression (with a median around 0.5), and that in some situations the cautious use of whole blood gene expression data could be a useful proxy measure of CNS gene expression.

To investigate the feasibility of large scale expression data collection, a pilot project called the Genotype-Tissue Expression project (GTEx; http://nihroadmap.nih.gov/GTEx/) was recently announced. The aim of this project is to develop a database containing expression data from approximately 1000 donor individuals in 30 different types of tissue. These individuals will also be genotyped at high density. It is hoped that with these data a comprehensive database of human eQTLs can be developed.

**PATHWAY ANALYSIS**

Variation or disruptions in different genes can have similar phenotypic consequences if the genes are involved in the same pathway. Disruptions at different stages of a pathway might all, independently or in interaction, lead to an increased risk of disease or expression of a complex trait. Pathway-analysis investigates whether a number of genes that have been found in e.g. a genetic
association study, are more often involved in a certain biological pathway than expected by chance (Wang et al., 2007).

Studies that have employed GWA and pathway analysis have reported some promising results (Ritchie, 2009). For example, Vink et al. (2009) searched for genes that may be involved in smoking behaviour, both initiation and persistence. Genes that showed an association with smoking behaviour in multiple samples were analyzed in terms of biological function, cellular location and possible interactions of the gene products. Using this approach they identified several groups of genes of similar function which may affect smoking behaviour. Several other phenotypes have been investigated using similar approaches, including multiple sclerosis (Baranzini et al., 2009), type 1 and 2 diabetes and bipolar disorder (Torkamani et al., 2008). Many others will undoubtedly follow. This type of analysis may be an important new step towards understanding the biological mechanisms underlying a trait.

Clearly, the introduction of genome-wide SNP arrays initiated many rapid developments in the field of gene-finding, and this may only be the beginning. New approaches such as the gene network and pathway-based analyses are only just starting to be developed. Although there have been many successes, there are also plenty of challenges left, especially in terms of the management and analysis of the huge amounts of data that are available already, and the even larger amounts of new data that are currently being collected, such as whole genome sequence data. Given the promising results published in recent years, we can only expect more to come.
References


Chapter 2


