Statistical Methods in Genetic Epidemiology

38

Heike Bickeböller and Duncan C. Thomas

Contents

H. Bickeböller (✉)
Department of Genetic Epidemiology, University Medical School, Georg-August-University of Göttingen, Göttingen, Germany

D.C. Thomas
Department of Preventive Medicine, Division of Biostatistics, University of Southern California, Keck School of Medicine, Los Angeles, CA, USA

W. Ahrens, I. Pigeot (eds.) *Handbook of Epidemiology, 2nd edition*,
DOI 10.1007/978-0-387-09834-0_62,

38.1 Introduction

Genetic epidemiology combines the scientific disciplines of human genetics, epidemiology, and biostatistics and has close relationships with the fields of medicine, molecular genetics, and molecular epidemiology. The latter traditionally has been concerned more with the study of molecular markers of exposure, susceptibility, and disease (see chapter ▸Molecular Epidemiology of this handbook). The field is also a specialized subdiscipline of biometry and mathematical population genetics with major biometrical contributions to human genetics and the development of statistical methods including segregation, linkage, and association analysis; simulation methods; and computer algorithms. Rather than focusing on cells or molecules (as in molecular genetics) or on individual patients (as in clinical genetics), genetic epidemiology research is conducted using populations or large series of systematically collected families (Khoury et al. 1993).

Genetic epidemiology aims to detect the genetic origin of phenotypic variability in humans (Vogel 2000) and unravel genetic components that contribute to the development or the course of a disease (or more generally a *phenotype*, the observed trait), along with environmental or other risk factors that may modify the effects of genes. Thus, the International Society of Genetic Epidemiology (IGES 2012) describes the field as a marriage between the disciplines of genetics and epidemiology, emphasizing the need to join the fields. Whereas genetics tends to focus on the genotype-phenotype correlation neglecting the environment and epidemiology tends to focus on environmental and demographic factors, a full understanding of the etiology of complex traits can only be achieved by considering both explaining how genes are expressed in the presence of different environmental contexts and how genetic and environmental factors act together in shaping a phenotype.

In contrast to classical risk factor epidemiology, the three main complications in genetic epidemiology are dependencies, large data sets, and the use of indirect evidence. The structure of chromosomes and families or populations leads to major dependencies within the data, thus requiring customized models and tests. Modern technologies can yield millions of genotypes per subject for many thousands of subjects at an affordable cost, and even higher density sequencing platforms are now becoming available, along with a plethora of other data types (e.g., expression, proteomics) and repositories of biological knowledge (*ontologies*). In many studies, the disease-related functionally relevant DNA variant(s) in a gene are not directly observed, and hence the evidence on them is only indirectly given through correlated variant(s).

This chapter is solely devoted to methods dissecting the genotype-phenotype correlation with a binary phenotype (affected/unaffected). It does not specifically cover quantitative phenotypes, although many of the techniques discussed below can be applied to such problems. Section 38.2 presents an overview of major study designs and types of analysis. Section 38.3 introduces the most important genetic models. Sections 38.4–38.6 cover the three major types of analysis, segregation, linkage, and association analysis. Section 38.7 describes recent developments and looks to the future. We offer some general conclusions in Sect. 38.8. Detailed information on diseases used as examples in this chapter may be found in the standard reference of McKusick (1998) or its online version, Online Mendelian Inheritance in Man (OMIM 2012).

38.2 Study Types

Genetic epidemiology investigations are usually triggered by epidemiological studies that demonstrate a positive family history as a risk factor for disease, suggesting the existence of genetic or shared environmental factors. Often the goal of initial studies is to estimate the *relative risk for relatives of affected individuals* compared to the general population, such as λ_S in the case of siblings of affected individuals, in order to support a genetic hypothesis.

To further investigate familial aggregation, a *segregation analysis* may be carried out in pedigrees. This aims to determine whether a *major gene* is influencing a given phenotype in these families and if so to estimate the parameters of the underlying genetic model. All methods for segregation analysis are based on probability calculations for the observed phenotypes conditional on hypothetical genetic model parameters and on family structure, that is *genealogies*. Parameter estimation is often based on likelihood-ratio tests in order to select the most plausible model nested within a hypothetical general model. Sometimes family studies are also used solely (twin studies) or jointly with a major gene to estimate the *heritability* h^2 of a trait, that is, the proportion of the variance explained by (additional) genetic components. Hence, λ_S or h^2 are often used to indicate the genetic basis of a phenotype in a population (or enriched families) before marker studies are performed.

The primary cause of a *monogenic disease* such as cystic fibrosis is a mutation within a single gene that segregates according to Mendelian laws (see below). The predisposing variants (the alleles carrying the risk) of this major gene are usually rare in the population. For *complex* or *multifactorial diseases*, there may still be Mendelian subforms such as breast cancer caused by the major gene *BRCA1*. For rare monogenic diseases and rare Mendelian subforms of complex diseases, segregation analysis and subsequent further analyses perform well. However, complex diseases in general require more sophisticated methods of analysis. For example, in Alzheimer's disease, there are at least three major genes and several susceptibility genes conferring moderate risk (*oligogenes*). Oligogenes can have relatively common alleles carrying the risk. *Polygenic* effects at many loci across the whole genome, each with a minor effect, may contribute to disease.

If there is evidence for the existence of genetic factors contributing to a disease, the next step is to identify susceptibility genes in order to quantify the genetic influence and to understand the underlying genetic model and pathway to the phenotype. To this end, measures of correlation between a *genetic marker* and the (unknown) *disease locus* are used. A genetic marker is a DNA segment for which the chromosomal localization is known and multiple alleles can be determined. In general, methods assume Mendelian segregation of the marker (see Sect. 38.3.2).

Frequently used markers are multiallelic *microsatellites* and biallelic *single nucleotide polymorphisms (SNPs)*. A marker is termed a *polymorphism* if the frequency of the most common variant is less than 99%.

Two types of correlation between a genetic marker and the susceptibility locus are used:

- *Linkage (cosegregation at the family level)*: Linkage is present if the transmissions of DNA at marker and disease susceptibility loci from a parent to a child are not independent. Relatives with a similar disease status (e.g., both affected) are then more similar at a marker close to the disease susceptibility locus than expected under independence.
- *Linkage disequilibrium (LD, association at the population level)*: LD is present if in a gamete the joint probability for a specific marker allele and a specific disease allele differs from the product of individual probabilities. In affected individuals, certain marker alleles will then be more or less frequent than in randomly selected individuals from the population.

Linkage analysis in families is based on linkage; *association analysis* in populations or families uses linkage disequilibrium. Some designs and corresponding statistical methods are capable of integrating both types of information into the analysis.

For the analysis of complex diseases with genetic markers, we can distinguish two major approaches: A *candidate gene investigation* focuses on genes (or genomic regions) whose function in the pathway to the phenotype is thought to be known. A prominent example of a candidate gene system is the HLA (human leukocyte antigen) complex on chromosome 6. HLA is involved in immune resistance and is thus a natural candidate gene region for all autoimmune diseases. The genotypes of the relevant functional component of the candidate genes are not always observed. In this case, we use the information on genetic markers that lie in or in close proximity to the candidate gene in question. In contrast, a *genome scan* – a systematic coarse grid search of the whole genome with genetic markers – aims to localize one or more regions harboring susceptibility genes. A typical scan might investigate approximately 350–700 microsatellites with an average distance of 5–10 cM (centiMorgan, see Sect. 38.3.3) or 500,000 or more SNPs along the genome, depending on the type of genetic information and design used (linkage or association, respectively).

38.3 Genetic Models

38.3.1 Terminology

The *genome* is the complete collection of an individual's genetic material present in every cell, consisting of *chromosomes* (long DNA strands). A *gene* is a piece of a chromosome coding for a function that can be seen as the heritable unit. The *locus* is the position of a piece along the chromosome. The locus might denote the position of, for example, a gene, a gene complex, or a marker. The different variants of a gene are called *alleles*.

The human genome is *diploid*, that is, chromosomes are paired *(homologous chromosomes)* with the exception of the sex chromosomes in males. Each human somatic cell contains 22 *autosomal* pairs and 1 pair of sex chromosomes. In a pair, the autosomal chromosomes contain the same gene with possibly different alleles at the same location. During *meiosis*, a diploid chromosome set is reduced to a *haploid* chromosome set of a germ cell, the *gamete*.

A pair of an individual's alleles at a locus is called a g*enotype*. If the alleles are identical, the individual is called *homozygous* at the locus, otherwise *heterozygous*. Two copies of a gene are called *identical by descent (IBD)* if both copies are identical and are copies of the same gene in a common ancestor. An individual is *homozygous by descent (HBD)* when its gene pair is IBD. When considering several loci simultaneously, the multilocus alleles inherited from the same parent constitute a *haplotype*.

38.3.2 Mendelian Single-Locus Model

Mendelian segregation (Mendel 1865) is the simplest and most commonly used model for the *mode of inheritance* for a single locus. An individual randomly and independently inherits one allele from father and mother, respectively. All segregation events from parents to offspring are independent. This implies that copies of some alleles are frequently present in offspring and other alleles are lost in subsequent generations, hence leading to random changes in population allele frequencies over time (*genetic drift*).

Consider the phenotype affected/unaffected by a certain disease. Let S denote a susceptibility gene with n alleles $S_1,\ S_2,\ldots,S_n$. The distribution of allele frequencies in the population is denoted by $p_r = P(S_r)$ $(r = 1,\ldots,n)$. Under Hardy-Weinberg equilibrium (HWE) (Hardy 1908; Weinberg 1980), the (unordered) genotype frequencies are given by

$$\begin{array}{ll} P(S_r S_s) = p_r p_s = p_r^2 & \text{for } r = s \\ P(S_r S_s) = 2 p_r p_s & \text{for } r \neq s \end{array}.$$

These frequencies follow from independence of the corresponding allele frequencies, combining two ordered genotypes for heterozygotes. Its maintenance in a population can be derived by applying Mendelian segregation to each possible parental mating type, assuming random mating (Khoury et al. 1993).

Penetrance describes the relation between genotype and phenotype. It is the conditional probability that an individual will be affected given its genotype: $f_{rs} = P(\text{affected} \mid S_r S_s)$.

Classical monogenic diseases are those caused by a single major gene, for which the penetrances take only the values 0 or 1. Often a locus S is assumed to be *biallelic*, that is, to have only two different alleles. Let S_1 denote the "susceptibility" allele (mutation) and S_2 the "normal" allele (wild type). For a classical *dominant* *disease,* all carriers of the susceptibility allele are affected ($f_{11} = f_{12} = f_{21} = 1$, $f_{22} = 0$); for a classical *recessive disease*, only homozygous carriers are affected ($f_{11} = 1$, $f_{12} = f_{21} = f_{22} = 0$).

Many classical hereditary diseases follow a Mendelian mode of inheritance. Often, the prevalence is below 1 in 1,000 live births. Examples are cystic fibrosis (autosomal recessive gene, cystic fibrosis transmembrane regulator (*CFTR*)) and Huntington chorea (autosomal dominant gene, *huntingtin* (HTT)). Both diseases can be caused by any of many different mutations. However, the assumption of a gene with a normal and a susceptibility allele (group) worked well in identifying these genes as disease causes, even though the true inheritance is much more complicated. The aim in statistical genetics is not to specify the correct and complete model but to address the scientific question adequately with a parsimonious mathematical model. If this model is too simple, then more complex or new biologically motivated models need to be considered.

The genotype-phenotype relation is complicated for many diseases. Individuals with a susceptible genotype can stay unaffected (*incomplete penetrance*) and those with a non-susceptible genotype can become affected (*phenocopies*). The penetrances at a specific gene locus may be different for different alleles and may depend on age, sex, environmental exposures, or other factors. For a general single locus with susceptibility allele S_1, we generally assume that $1 \geq f_{11} \geq f_{12} = f_{21} \geq f_{22} \geq 0$ and specifically for a recessive mode of inheritance that $f_{12} = f_{21} = f_{22}$ or for a dominant one that $f_{11} = f_{12} = f_{21}$. It is usually assumed that the parental origin of an allele has no influence on a disease, that is, $f_{12} = f_{21}$, although there is a growing literature on imprinting and parent-of-origin effects that violate this assumption (Zhou et al. 2010; Ainsworth et al. 2011).

38.3.3 Linkage

For the joint inheritance at two loci, independent Mendelian segregation does not generally hold, owing to crossover events and recombinations. *Gametes* (comprising one allele from each pair of chromosomes) are formed during meiosis, when homologous chromosomes are arranged next to each other and partly overlap. A chromosome breakage and a *crossover* – an exchange between homologous chromosomal segments – can occur. A *recombination* between loci A and B occurs when a gamete has a haplotype comprising a combination of alleles different from that on the same grandparental chromosome due to crossovers between the loci.

Consider the formation of gametes during meiosis displayed in Fig. 38.1. Between distant loci A and B (see Fig. 38.1a), a crossover is likely to result in a recombination of the haplotypes A_1B_1 and A_2B_2 to give the new haplotypes A_1B_2 and A_2B_1. If the two loci A and B are very close (see Fig. 38.1b), this is very unlikely. *Map distance* is defined as the expected number of crossovers between two loci (Haldane 1919). In expectation, the number of crossovers is roughly proportional to physical distance between two loci, so this distance measure

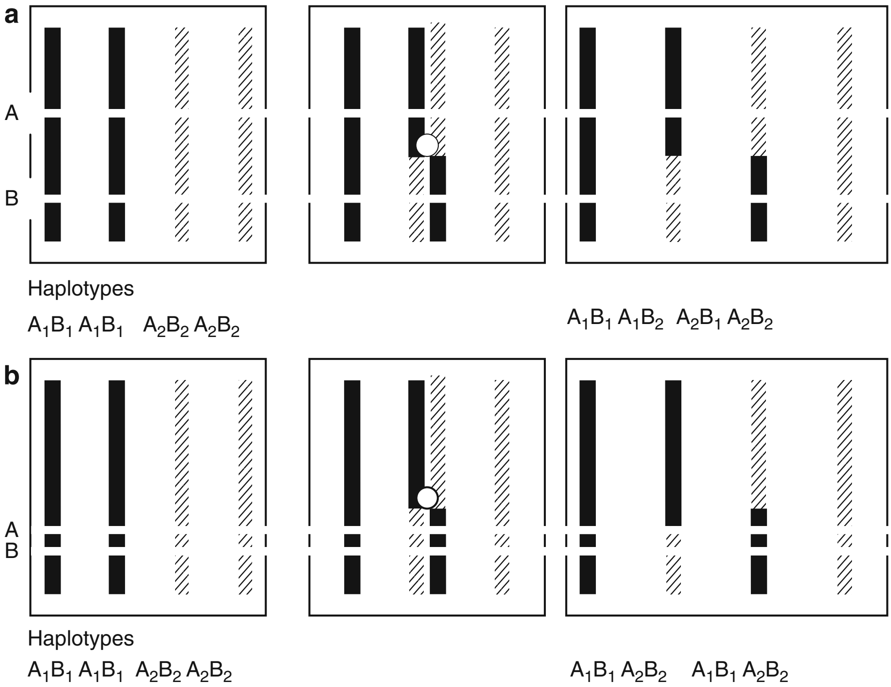
a
A
B
Haplotypes
A_1B_1 A_1B_1 A_2B_2 A_2B_2
A_1B_1 A_1B_2 A_2B_1 A_2B_2
b
A
B
Haplotypes
A_1B_1 A_1B_1 A_2B_2 A_2B_2
A_1B_1 A_2B_2 A_1B_1 A_2B_2

Fig. 38.1 Formation of gametes during meiosis from one parental pair of chromosomes with a single crossover. *Left*: parental chromosome pair, *middle*: crossover event (crossover point denoted by the *circle*), *right*: gametes for offspring formation. At the two loci A and B, the parent is double heterozygous A_1A_2 and B_1B_2. (**a**) The crossover occurred between locus A and B. The two middle gametes show recombination. (**b**) The crossover occurred above locus A and B so that the gametes do not show recombination

is additive, that is, for three (ordered) loci A, B, and C, the map distance between A and C is the sum of the distances between A and B and between B and C. The map unit is called a Morgan (M), named after T.H. Morgan, a Nobel prize winning geneticist (1866–1945), who discovered the importance of chromosomes for the inheritance process. The human genome contains approximately 3.3 billion base pairs with a total length of approximately 33 M, so as a rough guide, 1 centimorgan ($\mathrm{cM} = 0.01\,\mathrm{M}$) corresponds to one million base pairs in the physical map.

By genotyping, it is possible to observe recombinations between two loci, but crossovers are not directly observable. Figure 38.1a shows recombination due to a single crossover. For a double crossover (two chromosomal exchanges between loci), no recombination would be observed. This holds true for any even number of crossovers. Thus, a *recombination* is defined as an uneven number of crossovers between a pair of loci. The *recombination rate* θ – the ratio of the number of recombinant gametes to the total number of gametes formed – is used as a measure

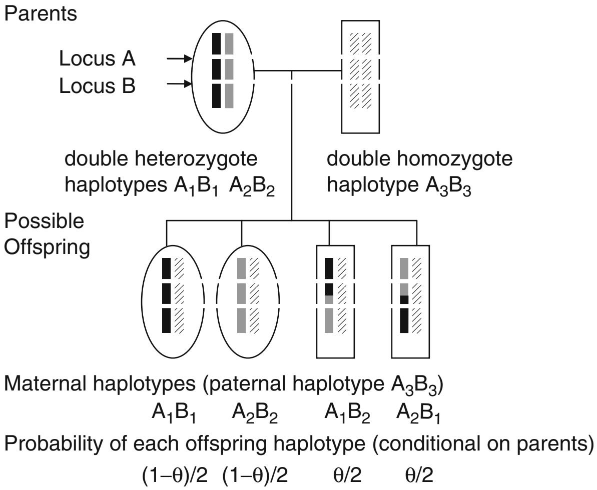
Parents
Locus A
Locus B
double heterozygote
haplotypes A_1B_1 A_2B_2
double homozygote
haplotype A_3B_3
Possible
Offspring
Maternal haplotypes (paternal haplotype A_3B_3)
A_1B_1 A_2B_2 A_1B_2 A_2B_1
Probability of each offspring haplotype (conditional on parents)
(1−θ)/2 (1−θ)/2 θ/2 θ/2

Fig. 38.2 Formation of recombinant and non-recombinant haplotypes by meiosis

of *genetic distance* between two loci. If loci are on different chromosomes or far away on the same chromosome, they segregate independently. By definition, there is *linkage* between the loci if $0 \leq \theta < 0.5$, and no linkage if $\theta = 0.5$. The closer loci are to each other, the less likely there will be crossovers and hence a recombination. Complete linkage (complete cosegregation) implies no recombination, and thus $\theta = 0$.

In Fig. 38.2, a double heterozygous parent with haplotypes A_1B_1 and A_2B_2 and a double homozygous parent with haplotype A_3B_3 are considered. For the double heterozygous parent, a meiosis can create the non-recombinant haplotypes A_1B_1 and A_2B_2 or the recombinant haplotypes A_1B_2 and A_2B_1. In order to determine recombination, a parent homozygous even at one locus is not informative. Given that recombination is present, each of the two recombinant haplotypes occurs with probability 0.5. Given that no recombination is present, each of the two non-recombinant haplotypes occurs with probability 0.5. For $\theta = 0.5$, there is independent segregation so that all four possible haplotypes are equally likely.

If three ordered close loci A, B, and C are considered, $\theta_{\mathrm{AC}} \approx \theta_{\mathrm{AB}} + \theta_{\mathrm{BC}}$. In contrast to the map distance in Morgans, recombination distances are not additive. *Mapping functions* provide a translation of recombination distances into map distance in Morgans. In the majority of chromosomal regions, recombination rates for women are higher than for men.

The potential informativeness of a single marker chosen from an existing marker map (without consideration of the disease locus) is determined by its genetic variability, that is, allele distribution. A measure of marker informativitiy is the *heterozygosity* H, defined as follows (Weiss 1993; Ott 1999):

$$H = \sum_{r \neq s}^{n} p_r p_s.$$

38.3.4 Linkage Disequilibrium

Linkage and linkage disequilibrium (LD) are different concepts. As linkage describes the coinheritance at two loci, it can only be observed in families, and it is independent of the specific alleles. LD describes the relation between alleles at two loci in a population.

Let S denote a locus with n alleles $S_1, S_2, \ldots, S_n$ and allele frequencies $p_r = P(S_r)$ and M a locus with m alleles $M_1, M_2, \ldots, M_m$ and allele frequencies $q_i = P(M_i)$. A common measure of LD is the haplotype probability minus its expectation under no association. For two biallelic loci, it is denoted by D or δ. For multiallelic markers, the parameter δ_{ir} is often used to define the linkage disequilibrium between M_i and S_r as

$$\delta_{ir} = P(M_i, S_r) - P(S_r)P(M_i), \qquad i = 1, \ldots, m;\ r = 1, \ldots, n.$$

LD is present if $\delta_{ir} \neq 0$ for any pair of alleles M_i and S_r. Under LD, the allele distribution at locus M is dependent of the S allele present. Often used measures of LD are D' (Devlin et al. 1996), defined as δ divided by its theoretical maximum for the observed allele frequencies, that is, a rescaling of δ to range between 0 and 1, and R^2, the square of the correlation coefficient.

LD can arise in several different ways (Suarez and Hampe 1994). At linked loci, complete LD can be caused by a recent mutation at one locus. However, disequilibrium is also possible without linkage between the loci (the term *gametic disequilibrium* is preferable in this case). One important mechanism for the development of disequilibrium at unlinked loci, even on different chromosomes, is *population stratification*. For example, through immigration or non-random mating (e.g., by religion or social status), populations may admix with different allele distributions in the populations.

Under random mating, LD decays over generations g according to $\delta_g = (1 - \theta)^g \delta_0$, where δ_0 is the initial LD at generation 0 (Maynard Smith 1989). Thus, whatever the origin of LD, in the presence of tight linkage, it can stay strong during many generations. Without tight linkage, LD will degrade rapidly. Thus, LD provides indirect evidence for linkage.

38.4 Segregation Analysis

The aim of *segregation analysis* is to test for the existence of a major gene influencing a phenotype and to estimate its mode of inheritance. The pattern of inheritance may be investigated in a few large families or in many small families.

Consider a Mendelian single-locus model for a major gene with susceptibility allele S_1 and normal allele S_2. In the classical Mendelian disease model, the penetrances $P(\text{affected} \mid \text{genotype})$ are only 0 or 1, so the genotype directly translates to a phenotype, and the families segregating the S_1 display characteristic disease patterns.

The simplest segregation tests (see, e.g., Sham (1998)) are based on *segregation ratios*, the proportion of affecteds among offspring of particular parental mating types. For illustration, consider a rare autosomal dominant disease and matings between an affected and an unaffected individual. These will usually be $S_1S_2 \times S_2S_2$ *matings*. Let r be the observed number of affecteds among n offspring and q the probability for a child to be affected. Then the segregation ratio is the unknown parameter q of a binomial distribution with sample size n. If the null hypothesis $q = 0.5$ is not rejected, it may be concluded that the data are compatible with an autosomal dominant disease pattern.

For each of six possible mating types ($S_1S_1 \times S_1S_1, S_1S_1 \times S_1S_2, S_1S_1 \times S_2S_2, S_1S_2 \times S_1S_2, S_1S_2 \times S_2S_2, S_2S_2 \times S_2S_2$), the distribution of genotypes and phenotypes in the offspring is determined by various genetic models. However, families are often recruited non-randomly according to particular ascertainment criteria, such as enrichment for disease, yielding an oversampling for particular parental genotypes, so for a valid test, the probability distribution needs to be corrected for this *ascertainment bias*. For example, if all families with at least one affected offspring are recruited (*truncate ascertainment*), the distribution of the number of affected offspring can be corrected for ascertainment by considering a truncated binomial distribution conditioning on $r \geq 1$ per family. If instead each case has an equal probability of being ascertained (*single ascertainment*), then multiple case families are represented proportional to the number affected and simply excluding the proband from the analysis may suffice. In practice, ascertainment schemes may be complex or unsystematic (Elston 1995), and misspecification of ascertainment might cause serious bias in the estimation of genetic parameters (see, e.g., Shute and Ewens (1988)).

For an *extended pedigree* with N individuals, a numerical procedure is needed. Let L denote the likelihood for the observed vector of phenotypes $Y = (Y_1, \ldots, Y_N)$, given a genetic model and the pedigree structure. L can be calculated by summing over all possible genotype vectors $G = (G_1, \ldots, G_N)_i$, $i = 1, \ldots, N$, in a given family, a particular one denoted by $g = (g_1, \ldots, g_N)$. We assume that the phenotype Y_i only depends on genotype G_i of that individual i. Thus, we get

$$L\,(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} P(Y|G = g)P(G = g).$$

The Elston-Stuart algorithm (Elston and Stewart 1971) provides an efficient recursive formula

$$L(Y) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_N} \prod_{j=1}^{N} P(Y_j|g_j) \prod_{k=1}^{N_1} P(g_k) \prod_{m=1}^{N_2} \tau\,(g_m|g_{m1}g_{m2})$$

where N_1 denotes the number of *founders* (individuals without specified parents in the pedigree, i.e., members of the oldest generation and married-in spouses) and

N_2 the number of *non-founders*. The parameters of the genetic model are (1) the genotype distribution $P(g_k)$, $k = 1, \ldots, N_1$, for the founders in the population for which HWE is usually assumed, (2) the transmission probabilities from parents m_1 and m_2 to offspring $m, \tau(g_m|g_{m_1}, g_{m_2})$, according to Mendelian segregation ($\tau(S_1|S_1S_1) = 1$; $\tau(S_1|S_1S_2) = 0.5$; $\tau(S_1|S_2S_2) = 0$), and (3) the penetrances $P(Y_j|g_j)$ relating genotypes to disease. This recursive formula works well on a *simple pedigree* of arbitrary size. As several unrelated pedigrees are independent from each other, their likelihoods can be multiplied to yield the total likelihood of a sample of pedigrees. Computations on *complex pedigrees* with marriage chains and inbreeding loops (such as consanguineous marriages) are often only possible with approximation methods.

Segregation analysis works well for monogenic diseases. Due to the unclear genotype-phenotype relationship, they are much more difficult in complex diseases. Several genetic and non-genetic factors, such as age, sex, and exposure factors, may have an influence on disease. Genetic heterogeneity may be caused by different alleles of the same gene or different genes or modifying factors that lead to different phenotypes segregating within a family (Evans and Harris 1992). In addition, there are further types of genetic heterogeneity such as genomic imprinting. In the presence of heterogeneity, considering homogeneous subgroups (defined by, say, severity, age of onset, family history, ethnicity) can lead to a clearer genotype-phenotype relation and thus to identify a possibly Mendelian subform of the disease.

An example of a highly successful segregation analysis for a complex disease is breast cancer (Newman et al. 1988). The families were ascertained through a population-based cancer registry. The ascertainment criteria for index cases were women with breast cancer, Caucasian, diagnosed before the age of 55 during a specified period with a histologically confirmed primary tumor. There was no selection on family history. Thousand five hundred and seventy-nine nuclear families were recruited, along with one large extended pedigree. Complex segregation analysis was applied using the program POINTER (Lalouel et al. 1983). It models an underlying unobserved quantitative trait called *liability* as a mixture of three normal distributions with different means for each of the genotypes, allowing for polygenes and an environmental component, with disease corresponding to the liability exceeding a certain threshold (Morton and MacLean 1974). For a predisposing genotype, the mean liability is shifted compared to the mean for non-disposing genotypes such that more individuals will exceed the threshold. Evaluation by direct modeling of the transmission probabilities allows the identification of the major factor as a major Mendelian gene. Population-based liability classes were taken from cumulative incidence rates estimated from a large epidemiological study in the region.

Segregation analysis is based on likelihood-ratio tests comparing different models. First, one investigates the consistency of the data with a major gene model; next, one considers which mode of inheritance fits the data best. To avoid false-positive results, the likelihood for Mendelian transmission probabilities can be compared against more general models corresponding to environmental or cultural transmission. For breast cancer, an autosomal dominant major gene provided the best fit, although the general single-locus model with three penetrances resulted in a comparable fit, and the non-Mendelian transmission models were strongly rejected.

Segregation analysis for a disease without a single major gene but only a few oligogenes may not be particularly rewarding. Many genetic marker studies are nowadays carried out without specification of the genetic model, but it is still worth establishing that a disease has some genetic basis by estimating the heritability h^2 before embarking extensive marker studies.

38.5 Linkage Analysis

In *linkage analysis*, the cosegregation between marker and disease within families is investigated to find evidence for linkage and often to estimate the recombination rate θ. The classical *lod score method* (Morton 1955) is a test for linkage between a susceptibility gene and a marker (null hypothesis H_0: $\theta = 0.5$ vs. alternative H_1: $\theta < 0.5$) under a parametric model for the genetic effect, allowing estimation of θ. For a detailed description, see Ott (1999). Let $L(\theta)$ denote the likelihood for the observed phenotypes at a particular value for θ conditional on the (assumed) model, the marker allele distribution, and the given pedigrees. In the usual notation, the underlying conditioning is sometimes left out. The lod score function ("log odds") is the log likelihood ratio

$$Z(\theta) = \mathrm{LOD}\,(\theta) = \log_{10} \frac{L\,(\theta)}{L\,(0.5)}$$

as a function of θ. $Z(\theta)$ compares the likelihood under linkage with recombination rate θ with the likelihood under no linkage, that is, $\theta = 0.5$. $Z(\theta)$ will be maximized over all possible values for θ, that is, $0 \leq \theta \leq 0.5$, to yield $Z_{\max}$. Values of $Z_{\max} > 3$ are taken as evidence for linkage. The recombination rate is estimated by $\theta_{\max}$, the θ-value corresponding to $Z_{\max}$. If $Z_{\max} < -2$, linkage can be excluded. The limits 3 and -2 are based on a sequential Wald test, such that the posterior probability for linkage when rejecting H_0 is 95% for a single alternative θ. As logarithms of base 10 are used, the limits correspond to stopping limits of 1,000 and 0.01 in the sequential testing procedure yielded by setting the probabilities of types I and II errors at $\alpha = 0.001$ and $\beta = 0.01$ (Morton 1955).

The likelihood $L(\theta)$ for linkage between two loci A and B for a sibship can be derived easily when the genotypes at both loci are directly observable. Let the mothers' genotypes be A_1A_2, B_1B_2 and the fathers' A_1A_1, B_1B_1; here, only the double heterozygous mother is informative. Without knowing the maternal grandparent's genotypes, one cannot determine the *phase*, that is, whether the mother's haplotypes are A_1B_1 and A_2B_2 (phase I) or A_1B_2 and A_2B_1 (phase II). If the phase were known, $L(\theta)$ would be given by a binomial distribution with

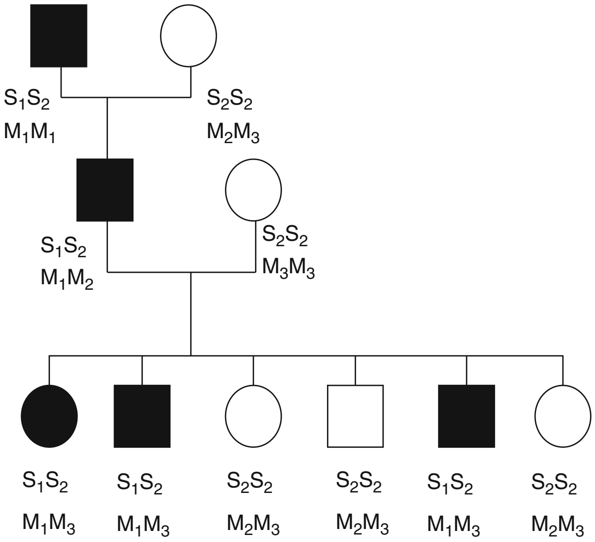

Fig. 38.3 Pedigree with a sibship of size six with marker information and with genotype information concerning the susceptibility locus, owing to the clear-cut rare autosomal dominant mode of inheritance (individuals: square = male, circle = female, black = affected, white = unaffected)

parameters n and θ, where n is the number of informative meioses. For an unknown phase, consider first phase I: Let n_x and n_y denote the number of meioses from the mother to the n children, of which n_x are non-recombinants ($\mathrm{A_1B_1}$ or $\mathrm{A_2B_2}$) and n_y are recombinants ($\mathrm{A_1B_2}$ or $\mathrm{A_2B_1}$). Under phase II, let n_x and n_y denote instead the number of recombinants and non-recombinants. Assuming LD phases I and II are both equally likely, then

$$L(\theta) = \binom{n_x + n_y}{n_x}\left[\frac{1}{2}(1-\theta)^{n_x}\,\theta^{n_y} + \frac{1}{2}\theta(1-\theta)^{n_x}\,\theta^{n_y}\right].$$

For the sibship in Fig. 38.3, let us now determine the likelihood $L(\theta)$, the lod score function $Z(\theta)$, $Z_{\max}$, and $\theta_{\max}$. Assume an autosomal dominant gene S with a rare susceptibility allele S_1 and a normal allele S_2. Thus, the affected father and all affected siblings have genotype S_1S_2. The marker M has alleles M_1, M_2, and M_3. The mother is homozygous and uninformative for linkage. She will not be considered further.

As a result of the genotyped grandparents, the father's haplotypes are known: S_1M_1 and S_2M_2. Thus, the phase is known and the likelihood is

$$L(\theta) = \binom{6}{0}(1-\theta)^6\,\theta^0 = (1-\theta)^6.$$

The lod score function is

$$\begin{aligned} Z(\theta) = \text{LOD}\,(\theta) = \log_{10} \tfrac{L(\theta)}{L(0.5)} &= \log_{10} \tfrac{(1-\theta)^6}{(0.5)^6} \\ &= 6 \log_{10} (1-\theta) + 6 \log_{10} 2 \\ &= 6 \log_{10} (1-\theta) + C \end{aligned}$$

where C denotes a constant independent of θ. The maximum of the lod score function is $Z_{\max} = 1.8$ for $\theta_{\max} = 0$. This corresponds to complete linkage as supported by no observed recombinations.

Missing information on grandparental genotypes in Fig. 38.3 results in an unknown phase. Then the lod score function would be

$$\begin{aligned} Z(\theta) = \text{LOD}\,(\theta) &= \log_{10} \tfrac{L(\theta)}{L(0.5)} = \log_{10} \tfrac{0.5\theta^6 + 0.5(1-\theta)^6}{(0.5)^6} \\ &= \log_{10} \left(\theta^6 + (1-\theta)^6 \right) + 5 \log_{10} 2. \end{aligned}$$

In this case, the maximum of the lod score function is $Z_{\max} = 1.5$ for $\theta_{\max} = 0$. Due to the uncertain phase, the maximum lod score is reduced. However, the estimate for the recombination rate stays at $\theta = 0$.

In Fig. 38.3, assume now that the second affected child has the genotype $M_2 M_3$ (and the genotype $S_1 S_2$). With the phase as indicated in the figure, one recombination needs to be taken into account now. Thus,

$$\begin{aligned} Z(\theta) = \text{LOD}\,(\theta) &= \log_{10} \tfrac{L(\theta)}{L(0.5)} = \log_{10} \tfrac{6\theta(1-\theta)^5}{6(0.5)^6} \\ &= \log_{10} \theta + 5 \log_{10} (1-\theta) + 6 \log_{10} 2. \end{aligned}$$

With one recombination, the maximum of the lod score function is $Z_{\max} = 0.63$ for $\theta_{\max} = 1/6 = 0.17$. Now linkage is estimated as incomplete, and $Z_{\max}$ is markedly reduced.

If in Fig. 38.3 the genotypes of the father and his parents are unknown, the father's genotype can be inferred as either $M_1 M_1$ or $M_1 M_2$. If HWE can be assumed, the likelihood of the recombination rate $L(\theta)$ can be calculated as a function of the marker allele frequencies in offspring. A detailed calculation will show that in this case, a rare marker allele M_1 will result in a high lod score and a more common marker allele M_1 will result in a lower lod score.

For a larger pedigree, $L(\theta)$ can be computed by the Elston-Stuart algorithm (Elston and Stewart 1971) described earlier, where now the g_j are the joint genotypes formed by the two loci S and M and θ is part of the transmission probabilities for the formation of gametes as recombinants or non-recombinants. Across families, the segregation process is independent, so $L(\theta)$ is simply the product of the individual family-specific likelihoods.

The lod score method has been very successful in localizing major genes, such as *BRCA1* for breast cancer (Hall et al. 1990), which was facilitated by focusing on early-onset families for which the genetic relative risk (*RR*) is strong and by having available a good segregation model. However, in complex diseases, the mode of inheritance is usually unclear, leading to false-positive and false-negative results as well as biased estimation of θ (Risch 1991; Lander and Schork 1994; Terwilliger and Ott 1994; Ott 1999). Joint segregation and linkage analysis often leads to biased parameter estimation, particularly if families are not systematically ascertained or the segregation model is misspecified. In MOD-score analysis (Risch 1984; Clerget-Darpoux et al. 1986), the LOD score is maximized over θ and the parameters of a biallelic single-locus model, that is, allele frequency and three penetrances. *Non-parametric methods* or *model-free methods* have been developed to avoid assumptions about the underlying genetic model. Their aim is to provide evidence for linkage without specifying the parameters of the underlying mode of inheritance and without estimating the recombination rate (Lander and Schork 1994; Elston 1998). They are often based on the *identity-by-descent (IBD)* status (Penrose 1953). For example, the IBD status of a patient and one of his/her siblings can take on the values 0, 1, or 2, according to the number of marker alleles that have been transmitted to both siblings from exactly the same grandparental copy of a parent's gene and are thus identical.

Allele sharing methods test whether relatives with a similar disease status (e.g., both affected) are more frequently similar in IBD at the marker than expected in the absence of linkage. In the *affected-sib-pair (ASP) method* (Day and Simons 1976), the observed counts of n ASPs with 0, 1, or 2 marker alleles IBD are compared with the expected ones assuming no linkage ($0.25n$, $0.5n$, or $0.25n$) using a χ^2-goodness-of-fit-test.

The literature on IBD methods is extensive and more powerful methods have been developed (e.g., Holmans 1993; Whittemore and Tu 1998). To determine IBD unambiguously, the marker must be sufficiently polymorphic, and the parents must be genotyped, or neighboring loci must be genotyped to yield sufficient information on the grandpaternal inheritance of the alleles. Often, IBD needs to be estimated. Sometimes, the *identity-by-state* (*IBS*) status (the number of identical marker alleles without considering ancestry) is used instead. In the Lander-Green algorithm for multipoint linkage analysis (Lander and Green 1987), the inheritance vector and thus the IBD status at a particular locus can be determined much more precisely even when some parents are not genotyped. The algorithm calculates the inheritance vector using a hidden Markov model walking from marker to marker, where pedigree size that a software can handle is limited by the length of this vector determined by the number of founders and non-founders in the pedigree (Lander and Green 1987; Kruglyak et al. 1996).

38.6 Association Analysis

The aim of *association studies* is to provide evidence for association or linkage disequilibrium in a population. LD results in an association between marker alleles and alleles of a susceptibility gene, such that certain marker alleles will be present more often in affected individuals than in a random sample of individuals from the population.

In classic *case-control studies,* marker allele frequencies or genotype frequencies in a group of unrelated affected individuals are compared to those in a group of unrelated unaffected individuals. Numerous associations have been identified with case-control studies, for example, associations of autoimmune diseases (e.g., diabetes, multiple sclerosis) with the HLA system or of apolipoprotein E (*APOE*) allele ε4 with Alzheimer's disease (Corder et al. 1993). The *APOE* ε4 allele frequency is approximately 35% in Alzheimer's patients, but only 15% in the older population not suffering from dementia. If a positively associated marker allele is frequent in a population, such as *APOE* ε4, then it is by itself not a good predictor for disease status, and the proportion of homozygotes for the allele is high. Linkage analysis methods are in general not very powerful in this situation.

Besides the usual limitations of classical case-control studies in epidemiology (cf. chapter ▸Case-Control Studies of this handbook), case-control studies to investigate linkage disequilibrium in genetic epidemiology must take a particular form of confounding known as population stratification into account. Population stratification denotes the presence of different ancestry populations, that is, discrete subpopulations or admixture of populations. If individuals are descended from populations with different allele frequencies and this is not taken into account, then spurious associations can be induced. To avoid this confounding, cases and controls must originate from the same homogeneous (including ethnically homogeneous) source population, or an appropriate design and analysis strategy needs to be employed.

If an association is found that is not considered spurious, this may have two causes (Lander and Schork 1994):

- The disease-associated allele is the susceptibility allele itself. If so, this association is expected to occur in all populations harboring this allele.
- The associated allele is in linkage disequilibrium with the susceptibility allele at the disease locus. If this is the case, then different associations can occur in different populations due to different haplotype frequencies at the two loci.

In the first case, marker and disease loci are identical, so $\theta = 0$, and LD is complete. In the second case, marker and disease locus are in general very close to each other. For this reason, association studies are highly valuable for the investigation of candidate genes.

As mentioned above, uncontrolled population stratification can result in spurious associations. For case-control studies, there are methods for taking the existence of subpopulations into account during statistical testing. All methods require many markers along the genome to be genotyped. In the *genomic control* method (Devlin and Roeder 1999), a variance inflation factor is used to adjust the test statistic, taking into account correlations between individuals in subpopulations. The *structured*

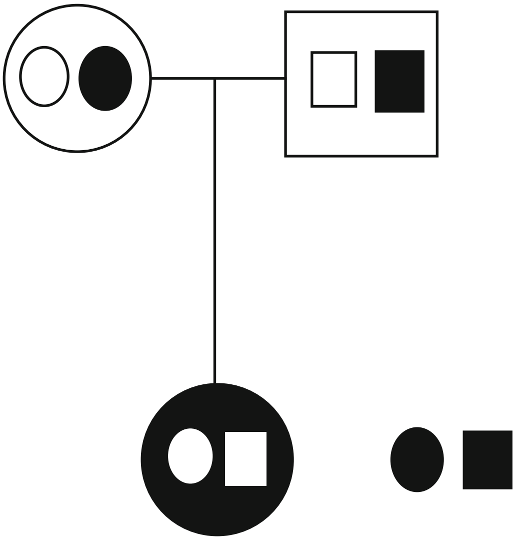

Fig. 38.4 Nuclear family with one affected child. Alleles transmitted from the parents to the affected child are denoted in *white*. Alleles not transmitted from the parents to the affected child are denoted in *black* (individuals: square = male, circle = female, black = affected, white = unaffected, inside: alleles)

association method (Pritchard et al. 2000) estimates the population structure and either assigns individuals to the most likely subgroup or better, estimates the proportion of each individuals' genome derived from each subgroup. Association is subsequently tested within subgroups or adjusted for ancestral source proportions. Both methods typically use panels of hundreds of "ancestrally informative" markers. *Principal component analysis* (Price et al. 2006) generally uses all the markers from a genome-wide scan and adjusts the association of any particular marker for the first few dozen principal components.

Family-based association studies avoid bias due to inadequate controls and population stratification by design. The concept of *internal controls* was first proposed by Falk and Rubinstein (1987). For the original design, nuclear families with at least one affected child are recruited, and the two parental alleles not transmitted to the affected child are used as internal controls (Fig. 38.4). With this design, information on both linkage and association between a marker and the susceptibility gene is used.

For a biallelic marker, the data resulting from this study design can be presented in various ways as 2 × 2 contingency tables and analyzed with standard statistical tests to investigate whether certain alleles are transmitted from the parents to an affected child more often than not (Terwilliger and Ott 1992; Schaid and Sommer 1994). Although in principle, all these procedures test for association (H_0: $\delta = 0$ vs. H_1: $\delta \neq 0$) and most for linkage as well, the most appropriate test respecting the matched nature of the transmitted and non-transmitted allele data is the McNemar test, which in this context is known as the *Transmission/Disequilibrium Test* (*TDT*)

Table 38.1 2 × 2 contingency table for family-based association studies based on a sample of N families with one affected child and both parents showing the matching of the two alleles of a parent. Consider a biallelic marker with alleles M_1, M_2. Small letters (a, b, c, d) denote allele counts. $2N$ denotes the total number of parental genotypes (i.e., of pairs of transmitted and non-transmitted alleles) to the affected child from the $2N$ parents

	Non-transmitted allele of one parent		
Transmitted allele of one parent	M_1	M_2	Total
M_1	a	b	$a + b$
M_2	c	d	$c + d$
Total	$a + c$	$b + d$	$2N$

(Spielman et al. 1993). The TDT is a haplotype-based analysis of the matched sample (Table 38.1). The test statistic is

$$\text{TDT} = (b - c)^2/(b + c).$$

The TDT compares whether the M_1 allele is more often transmitted to an affected child (b) than the M_2 allele (c) from heterozygous parents or vice versa. The test only considers M_1M_2 parents, since homozygous parents are not informative for preferential transmission of either allele.

The literature on family-based association analysis is vast (see, e.g., Whittaker and Morris (2001)). Important extensions of the above methods allow the application to multi-allelic markers, to tightly linked loci, and to quantitative traits. In addition, the design also allows for other types of nuclear families, such as sibships with affected and unaffected individuals (Spielman and Ewens 1998; Laird and Lange 2006). If a particular mode of inheritance is suspected, specialized versions of the TDT or related likelihood methods may yield higher power (Schaid 1999). If a candidate gene is to be investigated in detail, then a haplotype analysis can be carried out considering several biallelic polymorphisms (SNPs) in the same gene. The first step in a haplotype analysis is the estimation of the haplotype frequency in a population or the estimation of the most probable haplotype pair in an individual. For cohort or case-control studies, see Excoffier and Slatkin (1995), Stephens and Donnelly (2003), and Browning and Browning (2009); for family samples, see Rohde and Fuerst (2001) and Qian and Beckmann (2002). In the second step, linkage disequilibrium is investigated on the basis of the estimated haplotypes or haplotype frequencies. Some of these LD measures have already been described above (Devlin and Risch 1995).

Besides the analysis of main effects, gene-gene and gene-environment interactions can also be investigated in association analysis using standard tools. For gene-environment interaction, the case-only design enables the analysis of multiplicative interactions of factors, required to be independent in the population, on the basis of a sample of diseased individuals (Albert et al. 2001). If the independence assumption is valid, it is very efficient; if it is violated, such as for smoking-addiction genes and smoking, results can be severely distorted.

38.7 Current Methods and Outstanding Challenges

Genetics – and all its subdisciplines – has been an amazingly fast-moving field, with outstanding developments in both technology and biological insights. Methodological developments in statistical genetics and bioinformatics have had to scramble to keep pace; not the least of these challenges has been the daunting computational challenges posed by the massive data sets these advances have provided. In the remainder of this chapter, we review the current state of the art in the design of modern genetic epidemiology studies and the analysis of ultra-high-dimensional data and attempt to anticipate some of the novel developments that will be required in the foreseeable future.

38.7.1 Genome-Wide Association Studies

Fifteen years ago, Risch and Merikangas (1996) published a farsighted article in *Science* on the failure of traditional linkage analysis to uncover the genetics of complex diseases and made the then-radical suggestion that it would soon be possible to explore the entire genome by direct association methods. Their prediction came true, enabled by two developments in particular. The first was the advent of chip-based genotyping platforms that made it possible to assay hundreds of thousands to millions of SNP genotypes at a cost of under $1,000 per sample with high reliability. The second was a concerted effort by the public and private sectors to map the entire human genome (the Human Genome Project) and then to assemble a catalogue of known variants in a sample of Caucasian, African, and Asian populations (the International HapMap Project). In combination, these two advances provide a feasible way to directly genotype a large proportion of common variants and to predict many variants located next to the genotyped markers that are not typed directly.

It was nearly a decade before the first success of this approach was published in the form of a trio of papers on age-related macular degeneration in *Science*, one using this approach (Klein et al. 2005), which implicated a gene *CFH* in the complement pathway, a finding that has subsequently been confirmed numerous times. Since then, associations of about 210 diseases or quantitative traits and 1,300 genetic loci at genome-wide levels of significance ($p < 5 \times 10^{-8}$) have been published that have been independently replicated; the most recent version of the catalogue of published GWAS is available at the website of National Humane Genome Research Institute (NHGRI), National Institutes of Health (Hindorff et al. 2012; see also Hindorff et al. 2009). Study design and statistical analysis methods for such genome-wide association studies (GWAS) have been discussed in greater depth than is possible here (Hirschhorn and Daly 2005; Wang et al. 2005; Kraft and Cox 2008; McCarthy et al. 2008; Thomas 2010a, b; Witte 2010). Instead, we briefly review some of the recurring themes.

Multistage study designs. Early on, it was recognized that the multiple comparisons burden might be alleviated by some kind of staged design, in which only a portion of the sample would be used for screening the entire genome using one of the expensive high-density commercial platforms offering a fixed array of SNPs, followed by genotyping the remainder of the sample using a custom panel of only the most promising SNPs (Satagopan and Elston 2003). The final analysis then combined the information from both stages rather than treating it as a discovery and independent replication design (Skol et al. 2006); furthermore, it was possible to optimize the allocation of subjects to the two (or more) stages and the selection of the threshold for selecting SNPs to be genotyped in the later stages (Wang et al. 2006; Skol et al. 2007). With rapidly declining costs and increasing density of coverage of commercial panels, the interest in multistage designs has declined as investigators have recognized the advantage of having genome-wide data on the entire available sample as a resource for testing a broad range of hypotheses. However, the basic concept remains in spirit in the requirement for independent replication, as discussed in Sect. 38.7.3, and with the need for selecting a manageable number of individuals for next-generation sequencing technologies (Thomas et al. 2009).

Multiple testing and replication. With hundreds of thousands, if not millions, of associations being tested in a single study, there is an obvious need to avoid false-positive claims by adopting a stringent level of significance. A simple Bonferroni correction for one million SNPs would suggest a threshold of $0.05/10^6 = 5 \times 10^{-8}$, which has become the conventional criterion for claiming genome-wide significance, nominally ensuring a 5% probability of making at least one false-positive (family-wise error rate). This calculation fails to take into account the correlation among these tests due to linkage disequilibrium but is not a bad approximation for even the more recent platforms that allow testing of 2.5–5 million SNPs, as 1 million turns out to be roughly the effective number of independent tests in populations of European descent (or roughly double that in African-descent populations) (Pe'er et al. 2008). Fast asymptotic approximations have been developed (Conneely and Boehnke 2007) that allow for LD within a region, or permutation tests can be used as a gold standard for more complex dependency structures.

Despite the stringency of the genome-wide significance threshold, there are many factors than can lead to increased false-positives, some of which are discussed further below. But the key point is that a single study, however significant, is not usually considered convincing evidence of a genuine association without independent replication (Ioannidis 2007). Such replication is not simply to guard against chance variation (which can always be avoided simply by adopting an even more stringent significance threshold) but due to various sources of uncontrolled bias. Hence, real scientific replication should involve some elements of validation in different populations, by different investigators, using different methods. This is already demanded by the standard Bradford-Hill criteria in epidemiology used to establish validity of the association and strengthen belief in causality. This is not always possible in practice, however, such as for rare diseases where the discovery comes from a consortium of virtually all the available data or for studies conducted in unique settings (suggesting a need for some flexibility in the replication demands of granting agencies and journal editors!). Here in particular, the support by additional biological evidence is highly warranted.

Population substructure and study designs. One of the most pervasive sources of bias in GWAS is population substructure, as discussed in Sect. 38.6. Most GWAS therefore adjust instead for either an estimate of global ancestry from a finite set of founding populations using ancestry informative markers, typically using the STRUCTURE program (Pritchard et al. 2000) or using the top principal components from all or a subset of the markers, typically using the EIGENSTRAT program (Patterson et al. 2006). Either of these approaches tends to be quite effective at controlling the overall false-positive rate at least in homogeneous populations such as those of European descent. As a diagnostic for whether residual overdispersion due to uncontrolled population substructure remains after such adjustment, Quantile-Quantile plots of observed versus expected p-values of the single marker test statistics are generally used, and the genomic control overdispersion factor (Devlin and Roeder 1999) is checked to see if it is close to one. Control of population substructure can be more difficult in admixed populations like African-Americans and Hispanics, but these offer the advantage of being able to use within-individual comparisons for admixture mapping (Patterson et al. 2004; Freedman et al. 2006). For conventional GWAS scans in admixed populations, further adjustment for local ancestry (i.e., the ancestral origins of individual's chromosomes in the specific region of interest) may be necessary; the LAMP (Sankararaman et al. 2008) and HAPMIX (Price et al. 2009) programs can be used for this purpose.

Imputation. From the beginning, it has been understood that most associations discovered in a GWAS were unlikely to be directly causal, because only a small fraction of the genetic variation in the genome was being tested, but would hopefully reflect indirect associations with nearby causal variants through linkage disequilibrium between the causal and the marker loci. With the availability of the much more extensive catalogue of variation from the HapMap project, it has now become possible to infer the genotypes for most of the common variants in the genome by using imputation techniques from one of these standard reference panels (Li et al. 2010; Marchini and Howie 2010). Although programs such as MACH provide an assessment of the most likely genotype at each untyped locus, their use in association analysis would fail to account for the uncertainty in these imputations, essentially a form of measurement error leading to biased tests. A more appropriate procedure is therefore to use the estimated genotype probabilities or (under an additive model) the expected "gene dosage" as a continuous variable in a logistic regression analysis (Hu and Lin 2010).

Reporting. Given the potential range of problems and the approaches different investigators might take to addressing them, there is a need for some systematic guidance for how GWAS should be reported, if only to avoid subsequent problems in synthesizing the literature. Several authoritative statements have been issued by various groups to address this need, without imposing a straightjacket that would stifle investigators' creativity (Ehm et al. 2005; Chanock et al. 2007; Ioannidis et al. 2008; Hudson and Cooper 2009; Khoury et al. 2009; Little et al. 2009).

GWAS Summary. Over roughly the last 5 years, a consensus has emerged that most of the discovered novel associations of common, complex disorders with common SNPs have been relatively weak, with odds ratios typically in the range of about 1.2–1.6. Furthermore, even in the aggregate, these associations account for only a small portion of the total heritability estimated from classical twin or family studies (McCarthy and Hirschhorn 2008). While it is possible that these heritability estimates may be somewhat biased, it is certain that there remains a large portion of undiscovered genetic variation ("dark matter") to be accounted for (Hindorff et al. 2009; Manolio et al. 2009). Even for such a strongly related genetic and well-measured quantitative trait as height, the 180 loci that have so far been discovered based on meta-analysis of studies totaling over 183,000 individuals account for only about 10% of the total variability. Furthermore, it has been estimated that even with astronomical sample sizes, the number of loci of comparable effect sizes might rise to 600 but would still account for only about 20% of the heritability, which has been estimated at greater than 80% of the total phenotypic variation (Lango Allen et al. 2010). A variety of hypotheses have been advanced to account for this unexplained heritability, including rare variants, copy number variants, gene-environment and gene-gene interactions, and epigenetic effects, which will feature prominently as we move into the "post-GWAS" era.

38.7.2 Post-GWAS

Meta-analysis. Given the small effect sizes being sought and the enormous multiple comparisons penalty, most successful GWASs have required thousands of subjects. Nevertheless, as the experience with height demonstrates, no one study is likely to uncover more than a small fraction of the loci involved in a complex trait, and even larger sample sizes will be needed. Hence, the field has moved into a "Big Science" era, in which many investigators studying a given trait have formed consortia to pool all the available data for analyses of tens or hundreds of thousands of subjects (de Bakker et al. 2008; Zeggini and Ioannidis 2009). Some of these consortia have functioned simply to meet the replication requirement for each other's discoveries, but the more important purpose is to try to identify additional and weaker associations through a much larger sample size. This could in principle be accomplished by either a reanalysis of their combined raw data ("mega-analysis") or by meta-analysis of their summary statistics (Lin and Zeng 2010). In practice, the latter is usually much easier to accomplish, particularly if the different studies have used different genotyping platforms but can effectively impute genotypes for a larger, common set of HapMap SNPs (Zaitlen and Eskin 2010).

Fine mapping. Having identified one or more genome-wide significant and replicated regions, one might proceed to try to localize the region where a causal variant or variants might lie before attempting deep sequencing or functional studies. Here, the obvious strategy is simply to retest the available samples (or better, additional samples) with a higher density of markers in the region(s) of interest, but there are obvious trade-offs between sample size, number of regions that can be fine-mapped, region sizes, and density of markers (or criteria for selecting specific markers). These issues are amenable to methodological research, but there does not seem to be any generally agreed guidelines as of this writing. Given the speed the field is moving and the rapidly dropping costs of sequencing, many groups have decided to proceed directly to sequencing, bypassing the intermediate step of fine mapping.

Interactions and pathway analyses. Following an initial scan for main effects of SNPs (in either a single study or meta-analysis of several), much more remains to be explored. One possibility is that there could be larger gene-environment or gene-gene interaction effects that do not produce significant marginal effects. The obvious problem is that the number of possible interactions can be very much larger than the number of main effects: For a GWAS of one million SNPs, for example, there are half a trillion possible pairwise interactions. A simple Bonferroni correction for multiple comparisons would thus require a significance level of 1×10^{-13}, and of course, interaction tests would require much larger sample sizes than main effects even at the same significance level (Marchini et al. 2005). Power can be enhanced by "case-only" analyses based on an assumption of independence of the interacting factors in the source population. For example, a reanalysis of cleft palate data obtained substantially narrower confidence limits on the interaction between smoking and the *TGFα* gene (odds ratio $(OR) = 5.14$, 95% confidence interval (CI) = (1.68,15.7) for the case-only analysis compared with $OR = 6.57$, 95% CI = (1.72,20.0) for the case-control analysis), equivalent to a 30% reduction in sample size required for the same precision (Umbach and Weinberg 1997). Case-only analyses are, however, biased if this assumption is violated, as might arise due to LD among nearby pairs of loci, population stratification, or behavioral factors that induce an association between genes and environmental factors. To overcome these difficulty, various staged or hybrid approaches have been introduced (Evans et al. 2006; Kooperberg and Leblanc 2008; Mukherjee and Chatterjee 2008; Li and Conti 2009; Murcray et al. 2011). Although power will still be much lower than for main effects, these various methods generally yield much better power than a simple exhaustive search (Cornelis et al. 2012; Mukherjee et al. 2012). To make systematic study of gene-environment interactions of adequate sample sizes possible in the future, it is essential that investigators planning new GWAS design their studies to have appropriate environmental measurements and appropriate population-based sampling schemes. For example, the recent US National Institutes of Health (NIH) "post-GWAS" initiative aimed at synthesizing all the available data on five cancer sites, replicating findings, and characterizing genetic risks and their modification by environmental exposures has been limited by the fact that many of the available studies have not collected any environmental exposure data, and if collected, the choice of measurements was highly variable and ranged from very crude to very detailed assessment.

Another possibility is that there could be many SNPs that individually fail to be genome-wide significant but that jointly contribute to a common pathway. A variety of methods have been developed for identifying subsets of genes in known pathways that collectively are overrepresented among the top GWAS associations. Of these, the technique of Gene Set Enrichment Analysis (Wang et al. 2010), originally developed for gene expression data, has been most widely used. Hierarchical Bayes methods provide a more flexible regression-based approach to incorporate external knowledge about genomic, pathway, or functional annotation into the analysis of GWAS data (Lewinger et al. 2007). Cantor et al. (2010) provide an extensive review of these and other approaches to prioritizing GWAS associations for further investigation. Key to all these methods is the extent and quality of external databases such as the Kyoto Encyclopedia of Genes and Genomes and the Gene Ontology, which can be used for annotation in a systematic manner (Thomas et al. 2007). It is an interesting phenomenon that, after the failure of most candidate gene studies to yield replicable findings, the enthusiasm for the "agnostic" GWAS approach is now drifting back toward a synthesis of pathway-based and agnostic reasoning!

Functional studies. A somewhat unexpected finding from many GWAS is how few of the discovered associations lie in coding regions of genes (Hindorff et al. 2009). While some of the SNP associations could reflect LD with nearby coding variants that have not yet been discovered, it seems more likely that the majority will reflect variants in promoter regions of genes or long-range enhancers. Molecular techniques for functional characterization of causal associations will depend on the nature of the postulated effect, a topic which is beyond the scope of this article; for a recent set of recommendations, see Freedman et al. (2011). Nevertheless, there is a growing interest in "integrative genomics" approaches that can combine information across different types of data, such as SNPs, mutations in tumor tissues, transcriptomics, methylomics, metabolomics, and proteomics (Schadt et al. 2005; Hawkins et al. 2010).

38.7.3 Targeted, Whole-Exome, and Whole-Genome Sequencing

GWASs are based on an underlying "common disease, common variant" hypothesis (Reich and Lander 2001), which postulates that complex diseases are caused, at least in part, by common variants that can be effectively tagged by other SNPs in the region. The SNP panels used in most GWAS, based on one million or fewer SNPs, indeed are generally effective at tagging most "common" variants (conventionally defined as those with minor allele frequencies (MAF) of at least 5%). Newer generations of 2.5–5 million SNP panels will enhance the coverage of "uncommon" variants (those in the range of 1–5% MAF), but even these are not expected to provide good coverage of "rare" variants (less than 1% MAF). To discover these, direct sequencing will be necessary. The advent of several different massively parallel and fast "next-generation sequencing" (NGS) technologies (Davey et al. 2011) has now made this cost-effective, at least for targeted regions such as those around selected GWAS hits or the whole-exome. At the time of this writing, costs are typically around $1,000 per sample for whole exome sequencing and $5,000 for the whole genome, but the "$1,000 genome" (Mardis 2006) is anticipated in the near future. However, the data management and data analysis challenges are formidable. In addition to the storage problems of terabytes of raw data produced for even a single subject (petabytes for a typical study), NGS does not directly yield genotype calls but rather a random set of short sequence reads that must be aligned to cover the region of interest. The number of reads at any specific location is thus randomly distributed (with a mean given by the average depth of coverage), so the genotype can only be inferred probabilistically, taking into account possible errors in the reads themselves. Thus, in addition to the trade-offs mentioned earlier about region size and sample size, a further dimension to the design challenge is the necessary depth of sequencing, subject to a constraint on total cost (Sampson et al. 2011). The optimal design will depend upon the purpose – whether for discovery of rare variants or for testing association. One promising option is to perform the sequencing on pooled DNA samples (Futschik and Schlotterer 2010), which can considerably reduce costs, but this adds complexity to the optimization of the numbers of subjects and numbers of pools.

Multiple Rare Variant Analyses. One of the first analytical issues that arises concerns testing of association with rare variants. For feasible sample sizes, it is virtually impossible to test association with any single rare variant, both because of their rarity and the massive multiple comparisons penalty – for a typical whole-genome sequencing study, for example, one might expect to discover on the order of 20 million variants, most appearing only a few times. Interest has therefore tended to shift to tests of the "multiple rare variants" hypothesis (Price et al. 2010). The most commonly used technique is some form of "burden" test, which simply compares the total number of rare variants at a particular locus carried by cases and controls, possibly weighted in some fashion by their frequency or other characteristics (see Basu and Pan (2011) for a comparison of the available methods), but hierarchical Bayes methods that take account of model uncertainty offer a flexible and attractive alternative (Quintana et al. 2011).

Study designs for assessing causality. Most GWASs have used a case-control design with unrelated subjects for greatest statistical power. However, family-based designs offer two important advantages for studying rare variants. First, the sample can be enriched for rare variants by targeting cases with a strong family history. But more important is that looking at the pattern of cosegregation of variants with disease within families offers great potential for distinguishing causal variants from private polymorphisms that are simply circulating in the family but have nothing to do with the disease (Zhu et al. 2010; Shi and Rao 2011). This is essentially a form of linkage analysis, which has long been recognized as the design of choice for mapping rare major genes.

38.7.4 Risk Models and Translational Significance

Finally, as the number of genetic associations grows, it is natural to ask whether they can be used for genetic risk prediction. For genetic risk prediction, the purpose for the test must be distinguished, for example, a screening test for the general population or a high-risk population or a confirmation or exclusion test for a particular mutation for a particular family member of a disease-enriched family.

For major genes, segregation models as described above, using available genetic and non-genetic information and possibly indirect LD measures, can be used for prediction. For screening in populations, clearly a single SNP with relative risk less than 1.5 has very little value, but in the aggregate, one might consider a risk index based on all the known variants. For prostate cancer, for example, there are now more than 50 known GWAS associations, and a substantial portion of the population will carry several of these variants. Unfortunately, attempts to do this so far have not been particularly encouraging (Jostins and Barrett 2011; MacInnis et al. 2011; So et al. 2011; Newcombe et al. 2012). For prostate cancer in African-Americans, Haiman et al. (2011) reported a twofold gradient in predicted risk across quartiles of risk using 40 GWAS SNPs and 3.5-fold using only the 27 that were significantly associated among African-Americans, but the latter comparison may be subject to some overfitting. A clinically useful screening test should have both high sensitivity and high specificity (Kraft and Hunter 2009; Kraft et al. 2009; Janssens et al. 2011). Since both depend on where one draws the line between "normal" and "elevated" risk, a widely used measure of the overall performance of a screening test is the Area Under the Receiver Operating Curve (AUC) (Sanderson et al. 2005; Zou et al. 2007), obtained by plotting sensitivity against one minus specificity across the range of possible cut-points of the index (here, the predicted genetic risk score). For breast and prostate cancer, Machiela et al. (2011) found AUCs of 0.53 and 0.57, respectively, compared with 0.50 for an index that was no better than chance. Clinically useful risk indices would require an AUC of the order of 0.80 or better. For Crohn's disease, this appears attainable (Pharoah et al. 2002), but not at this point for most cancers (Chatterjee et al. 2011). Perhaps a more important question is what is the additional predictive value of genetic test results on top of established risk factors, including family history (Pencina et al. 2008; So et al. 2011). With the proliferation of direct-to-consumer genetic testing kits, some based on rather flimsy scientific evidence or with somewhat misleading advice about possible lifestyle changes to improve their risks, this question becomes of immediate translational significance (Hudson et al. 2007; Kaye 2008). See Levy et al. (2007) for a discussion of the scientific and ethical significance of the publication of the first complete human genome sequence. (The United States Genetic Information Non-discrimination Act (May 5, 2008) and a similar law in Germany (Gendiagnostikgesetz vom 31. Juli 2009 (BGBl. I S. 2529, 3672)) were enacted to protect individuals from discrimination based on genetic test results.) Although genetic risk prediction for the general population may still be some ways off, there is potentially greater utility in predicting genetic variation in response to treatments because these effects are likely to be much stronger due to the lack of time for evolution to eliminate deleterious variants, considering the recentness of most drug exposures (Altshuler et al. 2008).

38.8 Conclusions

The field of genetic epidemiology is in the midst of a fundamental paradigm shift. Originally based on methods for describing familial aggregation, testing for the existence of a genetic basis (segregation analysis), and localizing genetic causes (linkage analysis), the mainstay has become testing association with directly measured genotypes. A decade ago, this was feasible only for a modest number of variants in candidate genes, an approach that is now widely viewed as not having been particularly rewarding because of our lack of success in picking good candidates. With the advent of high-density genotyping platforms, agnostic scans for common variants across the entire genome have become popular and have led to many unexpected discoveries, albeit generally with rather small effect sizes that even in the aggregate account for only a modest proportion of the total estimated heritability of most complex diseases. These data are now being mined with sophisticated algorithms in the hope of identifying novel pathways across many of the suggestive, if not genome-wide significant, associations. Future directions are aimed at trying to identify the cause of the remaining unexplained heritability by targeted, whole-exome, or (soon) whole-genome sequencing – technologies that will pose formidable statistical and computational challenges – and by understanding the biological basis of the observed associations through regulatory, epigenetic, or other mechanisms.

Acknowledgements Dr. Bickeböller is supported by the German Research Foundation (DFG) grants (GRK 1644/1, KFO 241). Dr. Thomas is supported by NIH research grants ES019876, ES07048, MH084678, and HG005927.

References

Ainsworth HF, Unwin J, Jamison DL, Cordell HJ (2011) Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. Genet Epidemiol 35:19–45

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35:606–619

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86:6–22

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF Jr, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS (2007) Replicating genotype-phenotype associations. Nature 447:655–660

Chatterjee N, Park J-H, Caporaso N, Gail MH (2011) Predicting the future of genetic risk prediction. Cancer Epidemiol Biomark Prev 20:3–8

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet 81:1158–1168

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P (2012) Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol 175:191–202

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Day NE, Simons MJ (1976) Disease susceptibility genes – their identification by multiple case family studies. Tissue Antigens 8:109–119

de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 17:R122–R128

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–337

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Ehm MG, Nelson MR, Spurr NK (2005) Guidelines for conducting and reporting whole genome/large-scale association studies. Hum Mol Genet 14:2485–2488

Elston RC (1995) Twixt cup and lip: how intractable is the ascertainment problem? Am J Hum Genet 56:15–17

Elston RC (1998) Methods of linkage analysis – and the assumptions underlying them. Am J Hum Genet 63:931–934

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigrees. Hum Hered 21:523–542

Evans DG, Harris R (1992) Heterogeneity in genetic conditions. Q J Med 84:563–565

Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. PLoS Genet 2:e157

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, Oakley-Girvan I, Whittemore AS, Cooney KA, Ingles SA, Altshuler D, Henderson BE, Reich D (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci 103:14068–14073

Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG (2011) Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 43:513–518

Futschik A, Schlotterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218

Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, Rybicki BA, Isaacs WB, Ingles SA, Stanford JL, Diver WR, Witte JS, Chanock SJ, Kolb S, Signorello LB, Yamamura Y, Neslund-Dudas C, Thun MJ, Murphy A, Casey G, Sheng X, Wan P, Pooler LC, Monroe KR, Waters KM, Le Marchand L, Kolonel LN, Stram DO, Henderson BE (2011) Characterizing genetic risk at known prostate cancer susceptibility loci in African Americans. PLoS Genet 7:e1001387

Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–1689

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci 106:9362–9367

Hindorff LA, MacArthur J (European Bioinformatics Institute), Wise A, Junkins HA, Hall PN, Klemm AK, Manolio TA (2012) A catalog of published genome-wide association studies. Available at http://www.genome.gov/gwastudies. Accessed 8 July 2012

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common disease and complex traits. Nat Rev Genet 6:95–108

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Hu YJ, Lin DY (2010) Analysis of untyped SNPs: maximum likelihood and imputation methods. Genet Epidemiol 34:803–815

Hudson K, Javitt G, Burke W, Byers P (2007) ASHG statement* on direct-to-consumer genetic testing in the united states. Obstet Gynecol 110:1392–1395

Hudson TJ, Cooper DN (2009) STREGA: a ‘How-To’ guide for reporting genetic associations. Hum Genet 125:117–118

IGES (2012) International Society for Genetic Epidemiology. http://geneticepi.org/front. Accessed 8 July 2012

Ioannidis JP (2007) Non-replication and inconsistency in the genome-wide association setting. Hum Hered 64:203–213

Ioannidis JP, Boffetta P, Little J, O’Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD, Higgins JP, McCarthy MI, McDermott DH, Page GP, Rebbeck TR, Seminara D, Khoury MJ (2008) Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol 37:120–132

Janssens AC, Ioannidis JP, van Duijn CM, Little J, Khoury MJ (2011) Strengthening the reporting of genetic risk prediction studies: the grips statement. Genet Med 13:453–456

Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Hum Mol Genet 20:R182–188

Kaye J (2008) The regulation of direct-to-consumer genetic tests. Hum Mol Genet 17:R180–183

Khoury M, Beaty T, Cohen B (1993) Fundamentals of genetic epidemiology. Oxford University Press, Oxford

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP, Janssens AC, Ostell J, Owen RP, Pagon RA, Rebbeck TR, Rothman N, Bernstein JL, Burton PR, Campbell H, Chockalingam A, Furberg H, Little J, O’Brien TR, Seminara D, Vineis P, Winn DM, Yu W, Ioannidis JP (2009) Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 170:269–279

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor h polymorphism in age-related macular degeneration. Science 308:385–389

Kooperberg C, Leblanc M (2008) Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol 32:255–263

Kraft P, Cox DG (2008) Study designs for genome-wide association studies. Adv Genet 60:465–504

Kraft P, Hunter DJ (2009) Genetic risk prediction – are we there yet? N Engl J Med 360:1701–1702

Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S (2009) Beyond odds ratios – communicating disease risk based on genetic profiles. Nat Rev Genet 10:264–269

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Genet 7:385–394

Lalouel JM, Rao DC, Morton NE, Elston RC (1983) A unified model for complex segregation analysis. Am J Hum Genet 35:816–826

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci 84:2363–2367

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC (2007) The diploid genome sequence of an individual human. PLoS Biol 5:e254

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC (2007) Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol 31:871–882

Li D, Conti DV (2009) Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol 169:497–504

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34:816–834

Lin DY, Zeng D (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol 34:60–66

Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G, Grimshaw J, Scheet P, Gwinn M, Williamson RE, Zou GY, Hutchings K, Johnson CY, Tait V, Wiens M, Golding J, van Duijn C, McLaughlin J, Paterson A, Wells G, Fortier I, Freedman M, Zecevic M, King R, Infante-Rivard C, Stewart A, Birkett N (2009) Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. PLoS Med 6:e22

Machiela MJ, Chen C-Y, Chen C, Chanock SJ, Hunter DJ, Kraft P (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol 35(6):506–514

MacInnis RJ, Antoniou AC, Eeles RA, Severi G, Al Olama AA, McGuffog L, Kote-Jarai Z, Guy M, O'Brien LT, Hall AL, Wilkinson RA, Sawyer E, Ardern-Jones AT, Dearnaley DP, Horwich A, Khoo VS, Parker CC, Huddart RA, Van As N, McCredie MR, English DR, Giles GG, Hopper JL, Easton DF (2011) A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. Genet Epidemiol 35:549–556

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511

Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 37:413–417

Mardis ER (2006) Anticipating the 1,000 dollar genome. Genome Biol 7:112

Maynard Smith J (1989) Evolutionary genetics. Oxford University Press, Oxford

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156–165

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369

McKusick VA (1998) Mendelian inheritance in man. Catalogs of human genes and genetic disorders, 12th edn. Johns Hopkins University Press, Baltimore

Mendel GJ (1865) Versuche über Pflanzenhybriden. Verhandlungen des Naturforschenden Vereins, Brünn

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

Morton NE, MacLean CJ (1974) Analysis of family resemblance. 3. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503

Mukherjee B, Chatterjee N (2008) Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694

Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2012) Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. Am J Epidemiol 175:177–190

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ (2011) Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet Epidemiol 35:201–210

Newcombe PJ, Reck BH, Sun J, Platek GT, Verzilli C, Kader AK, Kim S-T, Hsu F-C, Zhang Z, Zheng SL, Mooser VE, Condreay LD, Spraggs CF, Whittaker JC, Rittmaster RS, Xu J (2012)

A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. Genet Epidemiol 36:71–83

Newman B, Austin MA, Lee M, King MC (1988) Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc Natl Acad Sci 85:3044–3048

OMIM (2012) Online Inheritance In Man. http://www.ncbi.nlm.nih.gov/omim/. Accessed 8 July 2012

Ott J (1999) Analysis of human genetic linkage, 3rd edn. Johns Hopkins, Baltimore

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32:381–385

Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172; discussion 207–212

Penrose LS (1953) The general purpose sibpair linkage test. Ann Eugen 18:120–124

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–181

Qian D, Beckmann L (2002) Minimum recombinant haplotyping in pedigrees. Am J Hum Genet 70:1434–1445

Quintana MA, Bernstein JL, Thomas DC, Conti DV (2011) Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet Epidemiol 35:638–649

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502–510

Risch N (1984) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363–386

Risch N (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1616–1617

Rohde K, Fuerst R (2001) Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. Hum Mutat 17:289–295

Sampson J, Jacobs K, Yeager M, Chanock S, Chatterjee N (2011) Efficient study design for next generation sequencing. Genet Epidemiol. doi:10.1002/gepi.20575 (Epub ahead of print)

Sanderson S, Zimmern R, Kroese M, Higgins J, Patch C, Emery J (2005) How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. Genet Med 7:495–500

Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303

Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717

Schaid DJ (1999) Likelihoods and TDT for the case-parents design. Genet Epidemiol 16:250–260

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies. Am J Hum Genet 55:402–409

Sham P (1998) Statistics in human genetics. Wiley, New York

Shi G, Rao DC (2011) Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genet Epidemiol 35:572–579

Shute NC, Ewens WJ (1988) A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am J Hum Genet 43:374–386

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31:776–788

So H-C, Kwan JSH, Cherny SS, Sham PC (2011) Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet 88:548–565

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Suarez BK, Hampe CL (1994) Linkage and association. Am J Hum Genet 54:554–559; author reply 60–63

Terwilliger JD, Ott J (1992) A haplotype based haplotype relative risk approach to detecting allelic associations. Hum Hered 42:337–346

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore

Thomas D (2010a) Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272

Thomas DC (2010b) Design and analysis issues in genome-wide association studies. In: Khoury MJ, Bedrosian S, Gwinn M, Khoury MJ, Bedrosian S, Gwinn M, Higgins JPT, Ioannidis JPA, Little J (eds) Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease, 2nd edn. Oxford University Press, Oxford

Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO (2009) Methodological issues in multistage genome-wide association studies. Stat Sci 24:414–429

Thomas PD, Mi H, Lewis S (2007) Ontology annotation: mapping genomic regions to biological function. Curr Opin Chem Biol 11:4–11

Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16:1731–1743

Vogel W (2000) Genetische Epidemiologie oder zur Spezifität von Subdisziplinen der Humangenetik. Med Genet 4:395–399

Wang H, Thomas DC, Pe'er I, Stram DO (2006) Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11:843–854

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Weinberg W (1980) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64:368–383

Weiss KM (1993) Genetic variation and human disease: principles and evolutionary approaches. Cambridge University Press, Cambridge

Whittaker JC, Morris AP (2001) Family-based tests of association and/or linkage. Ann Hum Genet 65:407–419

Whittemore AS, Tu IP (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228–1242

Witte JS (2010) Genome-wide association studies and beyond. Annu Rev Publ Health 31:9–20 4 p following 20

Zaitlen N, Eskin E (2010) Imputation aware meta-analysis of genome-wide association studies. Genet Epidemiol 34(6):537–542

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201

Zhou JY, Ding J, Fung WK, Lin S (2010) Detection of parent-of-origin effects using general pedigree data. Genet Epidemiol 34:151–158

Zhu X, Feng T, Li Y, Lu Q, Elston RC (2010) Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol 34:171–187

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115:654–657