

Analysis of epidemiologic studies of genetic effects and gene-environment interactions

Montserrat García-Closas, Kevin Jacobs, Peter Kraft, and Nilanjan Chatterjee

Summary

This chapter describes basic principles in study design, data analysis, and interpretation of epidemiological studies of genetic polymorphisms and disease risk, including the assessment of gene-environment interactions. The case-control design (hospital-based, population-based or nested within a prospective cohort) is frequently used to study common genetic variants and disease risk. Because of their widespread use, the analysis of case-control data will be the focus of this chapter. Two key considerations in the study design will be addressed: the selection of genetic markers to be evaluated, and sample size considerations to ensure adequate power to detect

associations with disease risk. Single nucleotide polymorphisms (SNPs) are the most frequent form of common genetic variation, thus the discussion on data analysis will be based on the evaluation of associations between SNPs and disease risk. This chapter will begin with the evaluation of quality control of genotyping data, which is a critical first step in the analysis of genetic data. A description of statistical methods will follow, aimed at the discovery of genetic susceptibility loci, including analysis of candidate SNPs and genome-wide association studies, haplotype analyses, and the evaluation of gene-gene and gene-environment interactions.

Introduction

The approaches to studying genetic susceptibility factors for disease have evolved very quickly over the last several years, due to advances in genotyping technologies, substantial reductions in genotyping costs, and improvements in the annotation of common genetic variation, particularly the most common type of variant, the single nucleotide polymorphism. These advances have enabled investigators to move beyond evaluating a few candidate variants in key genes, to conducting more comprehensive, as well as exploratory, evaluations of common genetic variation in candidate pathways/networks to disease, and performing genome-wide association studies (GWAS).

Over the last year, there has been an explosion of new discoveries of susceptibility loci for a wide range of diseases derived from GWAS (<http://www.genome.gov/gwastudies/>). This rapid trend of discoveries is likely to continue in the near future, as an increasing number of epidemiological studies use this approach to identify novel susceptibility loci. A major factor in the success of these breakthroughs has been the formation of very large collaborative efforts through consortia of studies that is creating unprecedented opportunities for discovery.

The discovery of disease susceptibility loci can bring about improvements in the understanding of disease etiology, and may ultimately lead to improvements in risk assessment, targeted preventive or screening strategies to reduce disease incidence and mortality, and improvements in therapy through the identification of drug targets. The aim of this chapter is to describe basic principles in study design and data analysis in studies on common genetic polymorphisms and disease. A discussion of biases, and other considerations in the interpretation of data analyses, is outside the scope of this chapter and can be found in previous publications (1–3).

Study design

The study designs used in molecular epidemiology studies, and a description of their advantages and disadvantages, can be found elsewhere in this book (see Chapter 14) and in previous publications (4). Discussed here are aspects of these epidemiologic study designs that are most relevant to studies of genetic susceptibility to disease. The heritability of a disease, or the proportion of variation in

disease susceptibility due to genetic factors, is directly related to the ability to identify susceptibility loci in epidemiological studies (5). Therefore, one of the first considerations is to evaluate the heritability, or *a priori* evidence, that the disease of interest is caused by genetic variation.

The case–control design (3,6), either nested in a prospective cohort or by retrospective sampling of a population, is by far the most commonly used design in genetic epidemiology studies of unrelated individuals. Hospital-based case–control studies are particularly popular, as the hospital setting facilitates the rapid enrolment of subjects, and the collection and processing of biological specimens with high participation rates. The case–control design is of particular importance for the study of uncommon diseases that occur in small numbers in the population or prospective cohort studies. Given that most members of a prospective cohort will not develop disease, these studies often use sampling strategies, such as nested case–control and (less commonly) case–cohort designs, to improve efficiency (7). In these designs, only samples from cases and a random subset of non-cases are analysed, reducing the DNA extraction and genotyping costs considerably. The case–cohort design allows for the evaluation of several disease endpoints using the same comparison group (referred to as a subcohort); however, since the same disease-free subjects are repeatedly used as controls for different disease endpoints, depletion of DNA samples from this group can be an issue. Until whole-genome sequencing is cost-effective and commonly available, whole-genome amplification of DNA, from cases and controls can be used to

address the problem of limited DNA in epidemiologic studies; however, this amplified DNA might not be suitable for all genomic assays (8).

Biased sampling, or non-random selection of cases and/or controls, can be used to improve efficiency to discover genetic markers associated with disease. For instance, selection of cases with a family history of breast cancer can lead to gains in power to detect genetic susceptibility loci, assuming a polygenic model of inheritance with loci interacting multiplicatively (9). However, the generalizability of risk estimates and evaluations of gene–environment (G–E) interactions can be compromised.

Genotyping hundreds of thousands of genetic markers in thousands of individuals can be costly. Multistage designs are commonly used to reduce the cost of genotyping very large numbers of samples (10). In these designs, a proportion of samples are genotyped for a large number of markers (e.g. SNPs that represent genetic variation across the genome in GWAS). In subsequent stages, only those markers showing the most significant associations with disease are genotyped in additional samples (10). The reduced cost is offset by a reduction in power compared to a study genotyping all markers in all available samples. Since the majority of genetic association studies use some sort of case–control design, the description of methods for data analyses in the section *Analysis of Genetic Data*, which follows, will focus on case–control data.

Selection of genetic markers

This chapter focuses on single nucleotide polymorphisms (SNPs) – the most common form of variation in the human genome. A

SNP is a DNA sequence variation occurring when a single nucleotide base differs among members of a population. There are thought to be at least 10 million SNPs in the human genome, and the vast majorities are bi-allelic, having only two alleles or nucleotide variant forms. SNPs occur throughout the genome and can be measured (genotyped) accurately. Although the genotyping costs have decreased dramatically in the last few years, it is still cost-prohibitive to genotype all known SNPs or sequence the entire human genome. Therefore, current studies must select subsets of markers to be evaluated.

SNP selection strategies take advantage of the correlation among genetic variants located close together on the same chromosome, or linkage disequilibrium (LD), to select a minimal set of tag SNPs that capture the majority of common genetic variation in human populations (11,12). The selection of tag SNPs has been aided by the International HapMap Project (<http://www.hapmap.org/>), a public resource that has genotyped millions of SNPs in 270 individuals from different ethnicities (30 Yoruba from Ibadan, Nigeria, 45 Japanese residents of Tokyo, 45 Han Chinese, and 30 Caucasian trios from Utah, USA) (13). Several methods have been proposed to use extensive data sets like the HapMap to select tag SNPs. Pairwise tagging is a method where tag SNPs are selected by examining the LD measures between pairs of SNPs using a squared correlation coefficient, r^2 . A SNP is said to 'cover' another SNP if the r^2 value between them exceeds a given threshold (e.g. 0.80). The Carlson algorithm to select optimal tag SNPs is iterative and begins by considering all SNPs as potential tags. At each step, the SNP that covers the most correlated SNPs is chosen as a tag SNP. That

SNP and all other SNPs that it covers (called a bin) are removed, and the algorithm begins again and continues until all SNPs are either taken as tags or are covered by a tag (14). Multimarker or aggressive tagging algorithms examine correlations among two or more SNPs using a generalized correlation coefficient to determine coverage (15,16). This approach typically reduces the total number of tag SNPs required; however, the selection algorithm is computationally more intense than pairwise methods, and statistically more complex, since an appropriate multimarker test should be used to test the associations with disease. Multimarker tagging approaches are also more affected by missing genotype data, since several SNPs are often required to perform tests. The current generation of genotyping arrays used to perform GWAS include about 300 000 to 1 million SNPs to capture common genetic variation. The proportion of SNPs in HapMap covered by SNPs in each of these genotyping assays depends on the ancestral origin of the underlying population.

Sample size considerations

As in any epidemiological study, sample size considerations are critical for the design of studies of genetic associations and G-E interactions (10,17,18). The main parameters that determine the required sample size to attain a specified statistical power are:

- Disease prevalence in the population
- Magnitude of association (often measured by the odds ratio)
- Alpha-level or P-value threshold to designate a 'statistically significant' finding
- Genotype or allele frequency in the population
- Mode of inheritance

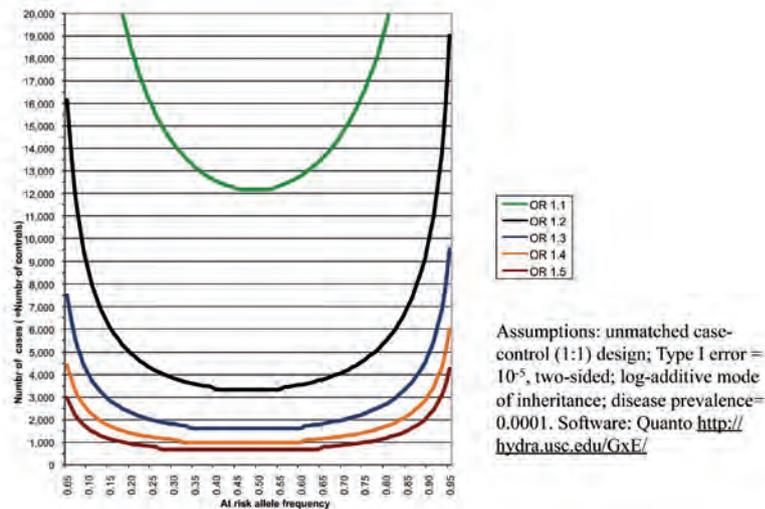
Generally, hundreds to thousands of subjects are needed to evaluate genetic associations with risk of complex diseases, as the magnitude of association between individual genetic variants on disease risk tends to be small (see <http://www.genome.gov/gwastudies/> for a catalogue of discoveries using GWAS in different diseases and traits).

Most studies measure genetic markers for disease rather than directly measuring the causal variant itself, as it is often unknown. The sample size needed to detect an association between a genetic marker and disease depends on the degree of linkage disequilibrium, or correlation due to physical proximity, between the marker and the causative variant. In approximate terms, the sample size requirements for studies using single SNPs as genetic markers are increased by a factor of $1/r^2$, where r^2 is the squared correlation between the marker and the causal unmeasured SNP (19). Sample size approximations are more complex when the disease susceptibility locus is in LD with multiple SNPs (10).

One limitation of standard power calculation methods is that they focus on the power for the detection of a single susceptibility locus with a given minor allele frequency (MAF) and disease odds ratio. In GWAS, however, there is likely to be a variety of susceptibility SNPs with a spectrum of MAF and disease odds ratios. The goal is to discover a certain number of underlying susceptibility loci, not some specific loci. A recent report has suggested novel approaches to power calculation that can provide realistic assessment for several probable discoveries in GWAS, accounting for the likely distribution of effect sizes for the underlying susceptibility loci (20).

An important challenge in large-scale evaluations of candidate genes/regions/pathways and GWAS, is to identify the few variants truly associated with disease among the large number being tested (21,22). Given the low probability of a true association (i.e. low prior probability) and the small expected magnitude of true associations (often resulting in low statistical power, particularly for less common variants), the standard threshold for statistical significance of an α -level of 0.05 results in the identification of a very high percentage of false-positive findings (23). Therefore, several authors have recommended reducing the P-value threshold to maintain a low probability that a statistically significant finding is a false-positive (i.e. false-positive report probability (FPRP)) (24,25). For instance, P-value thresholds of 10^{-4} – 10^{-5} were estimated for variants in candidate genes, and 10^{-7} for random variants to reach high probabilities of true findings ($\sim >80\%$) (25). Inversely, FPRP also depends on the prior and statistical power to detect an association. Therefore since the priors are often low, to reach a desirably low FPRP, the sample size of the study should be large enough to attain adequate statistical power. For instance, a P-value of 0.0024 for a SNP with a prior probably of 0.001 in a study of 300 cases and 300 controls will correspond to an FPRP of 72%; however, increasing the sample size to 1500 cases and 1500 controls, and keeping everything else constant, would lower the FPRP to 20% (23). Figure 16.1 shows the sample size requirements to detect genetic associations with a per-allele OR ranging from 1.1–1.5 for a variety of frequencies of the at-risk allele (assuming a log-additive mode of inheritance) and P-value threshold of 10^{-5} . Sample size

Figure 16.1. Sample size requirements to attain 80% power to detect a range of per-allele odds ratios (OR) for an association between disease risk and a bi-allelic SNP



needs increase dramatically for small changes in the OR when the magnitude of the OR is small, and for allele frequencies in the extremes (i.e. away from 0.50). The minimum sample size to detect a per-allele OR of 1.2 (i.e. assuming log-additive risk per allele and homozygous variant OR = 1.44) is 3300 cases and 3300 controls, whereas at least 12 000 cases and 12 000 controls are needed to detect a per-allele OR of 1.1 (i.e. homozygous variant OR = 1.21). Sample size needs would increase by a factor of 1.4 if a P-value threshold of 10^{-7} were to be used instead of 10^{-5} . These numbers illustrate that current studies of hundreds or a few thousand cases and controls have adequate power to detect an OR between 1.2–1.5 for common risk alleles (frequency $> 10\%$); however, much larger studies are needed to detect ORs of 1.1 for less-common risk alleles.

The statistical power of multistage GWAS designs depends on several factors: total number of available samples, number of samples and markers genotyped in each stage, α -level, the size of the

genetic effects to be detected, and type of analysis (10). The price of genotyping for different technologies used at each stage is also an important factor determining the optimal design of multistage studies. In general, joint analysis of data from the different stages is more powerful than replication analysis (26). As the cost of genotyping and sequencing methods continue to decrease, studies will be able to scan all individuals and eventually obtain a full genomic sequence, which will allow the evaluation of rarer variants as well as the mapping of causative variants. Sample size requirements for more complex analyses of genotype data, such as pathway-based, haplotype and novel high-dimensional analyses, are less well understood.

Evaluation of G-E interactions often requires large sample sizes that are further increased by the presence of errors measuring environmental and/or genetic exposures, even when the errors are small (17,27). Although multiplicative parameters for G-E interactions tend to be attenuated

by differential misclassification of exposure (17), this type of bias could lead to overestimation of the main effects of the exposure, joint effects, and subgroup effects, or additive interactions. Thus, high-quality exposure assessment and almost perfect genotype determinations are required for the evaluation of G-E interactions. This highlights the importance of validating genotype assays and including quality control samples during genotype determinations. This will help assess the reproducibility of the assays to identify problematic ones for possible re-genotyping or assay optimization.

There are several free statistical software programs to carry out power calculations for genetic association studies. POWER (<http://dceg.cancer.gov/bb/tools/power>) can be used for binary outcome studies (case-control or cohort studies) based on a logistic-like regression model with one or two covariates (e.g. gene-exposure interactions) (18); POWER for Genetic Association Analyses (<http://dceg.cancer.gov/bb/tools/pga>) can be used in case-control studies, fine-mapping studies, and whole-genome scans, for power and sample size calculations under various genetic models and statistical constraints; QUANTO (<http://hydra.usc.edu/gxe/>) is useful in matched case-control, case-sibling, case-parent, and case-only designs to compute sample size or power calculations to evaluate genetic associations, G-E interaction, or gene-gene (G-G) interaction; the CaTS Power Calculator (<http://www.sph.umich.edu/csg/abecasis/CaTS/>) is a user-friendly interface for power calculations for large genetic association studies, including two-stage GWAS (26); a spreadsheet can be downloaded to calculate

FPRP (<http://inci.oxfordjournals.org/cgi/content/full/96/6/434/DC1>) (23).

Current case-control or cohort studies usually include between a few hundred to a few thousand cases and a similar numbers of controls. Therefore, to meet the larger sample size requirements to identify weak associations (Cf. Figure 16.1) and interactions, especially when considering disease subtypes, an increasing number of consortia of existing studies have been and continue to be formed (28). Consortia can achieve the large sample sizes necessary to confirm or refute associations by coordinating the analysis of pooled data from many studies, as well as evaluating consistency of findings across studies of different quality and with different sources of biases (29). However, comparability of data on environmental exposures across studies may be a limitation. Therefore, very large, well-designed studies with high-quality exposure data and tumour specimens might be needed. To date, there are very few examples of gene-environment interactions that have been demonstrated in large pooling efforts. One example is the demonstration of interactions between cigarette smoking and polymorphism in the *NAT2* and *GSTM1* genes in the context of a bladder cancer GWAS (30).

Analysis of genetic data

Quality control of DNA and genetic data

Quality control analyses are conducted both before and after genotyping of DNA samples (31). Ideally, DNA samples should be accurately quantified before genotyping (e.g. using fluorescence nucleic stains, such as PicoGreen® (Molecular Probes Inc.)), and

profiled to obtain a “DNA fingerprint” using a panel of genetic markers that uniquely identify each sample (e.g. the Amp/STR® Identifiler® kit (Applied Biosystems) uses 15 SNPs and the Amelogenin marker for gender determination). This allows precise verification of duplicate DNA samples, identification of unexpected duplicates (e.g. due to sample collection, storage, labelling or plating errors), identification of gender mismatches between the DNA and self-reported gender, and identification of contaminated samples that should be excluded from further analyses. After genotyping assays have been performed, the quality of the resulting genotyping calls can be assessed by evaluating the scatter plots of allele-specific probe intensity values used for genotype determination. SNP genotype calls are made based on the clustering patterns of the probes, where clusters for each homozygote and heterozygote genotype state should be observed. High quality assays will demonstrate tight clusters with clear separation between them (Figure 16.2A versus 16.2B).

Genotyping completion rates can be calculated for DNA samples or loci:

- Overall completion rates—number of loci with genotype calls divided by the total number of genotyped loci
- Completion rates by sample—number of loci with genotype calls for a given sample divided by the total number of genotyped loci in that sample
- Genotype completion or call rates—number of samples with genotype calls for a given SNP divided by the total number of genotyped samples

Decreased completion rates often reflect poor assay performance, which may be due to

chemical and physical properties of the assay or the quality of the input DNA. Completion rates should also be calculated separately by DNA source, processing laboratory, DNA extraction method, case–control status and genotyping plate to detect systematic variation in genotype quality. Low genotype call rates can help detect loci with problematic assays that require re-genotyping, a new assay, or selection of a surrogate SNP. Completion rates by sample or plate can detect problems with specific samples or plates that could result in exclusion of data from those samples or plates. Analyses of completion rates by case–control status can detect assay performance differences due to varying DNA quality for cases and controls, which would result in differential misclassification. When large numbers of SNPs are genotyped, such as in GWAS, it is useful to look at the distribution and plot the completion rates by sample or loci (genotype calls) to detect outliers. Figure 16.3A shows an example of such plots, which utilize data from a scan using the Illumina HumanHap 1M assay with good

overall completion rates for most samples and loci.

Sample heterozygosity is the percentage of heterozygous genotypes in autosomal SNPs for a given sample. For instance, SNPs included in the Illumina HumanHap 1M genome-wide genotyping assay in populations of European origin have a mean heterozygosity of about 27%. Although samples from different racial origins will have different heterozygosity values, extreme outlier values can reflect sample quality or assay performance problems, which is reflected by a correlation between high (or low) sample heterozygosity and reduced sample completion rates. Plotting heterozygosity for all samples and against sample completion can help identify low performing samples (see Figure 16.3B for an example).

Analyses of data from duplicate quality control samples include calculation of percent agreement of informative genotypes (i.e. concordance of non-missing genotype calls for DNA samples from the same individual) among pairs of samples. As with completion rates, good genotype concordance should also

be evaluated by plate and sample, since this can give clues as to the source of error (e.g. systematic errors often reflect sample handling or plate labelling and orientation problems, whereas random errors reflect assays' reproducibility).

In very large, randomly mating populations with no selection, genotype frequencies are expected to be constant and 'in equilibrium' from generation to generation. This phenomenon is called Hardy–Weinberg equilibrium (HWE), and the expected genotype frequencies under HWE are called Hardy–Weinberg proportions (HWP). It should be noted that although HWE and HWP are often used interchangeably, HWE is a multigenerational phenomenon and cannot be directly assessed in standard epidemiological studies. Under random mating and no selection, HWP implies HWE; however, under selection and non-random mating, genotype frequencies can be in HWP but not HWE. For a bi-allelic SNP with A-allele frequency p and a-allele frequency $q = (1-p)$, the expected genotype proportions under HWP

Figure 16.2. Examples of genotype clustering plots used to make genotype calls (polar cluster plots of the normalized intensity and allelic intensity ratio)

A. Good assay with clear separation of genotype clusters

B. "Poor" assay with overlapping genotype clusters

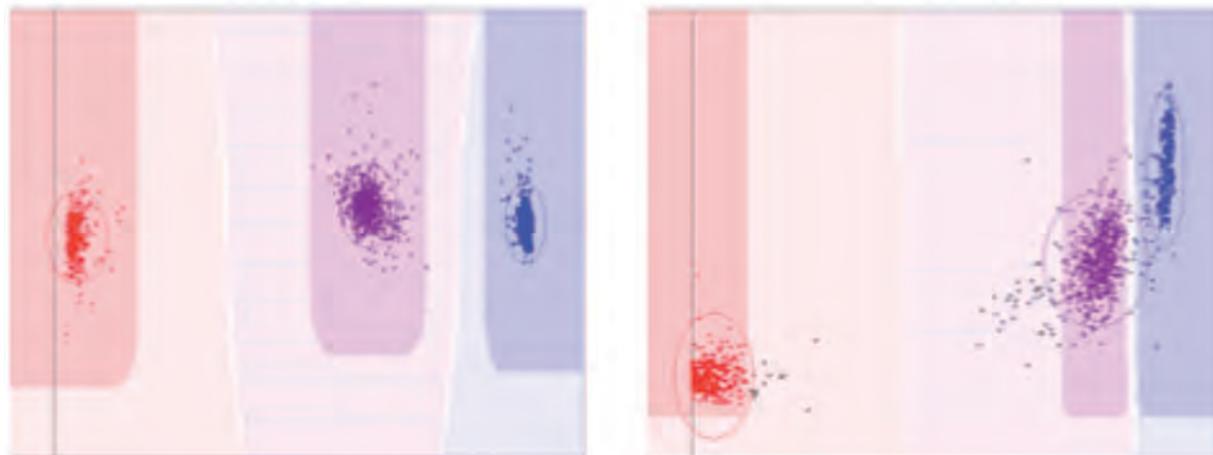
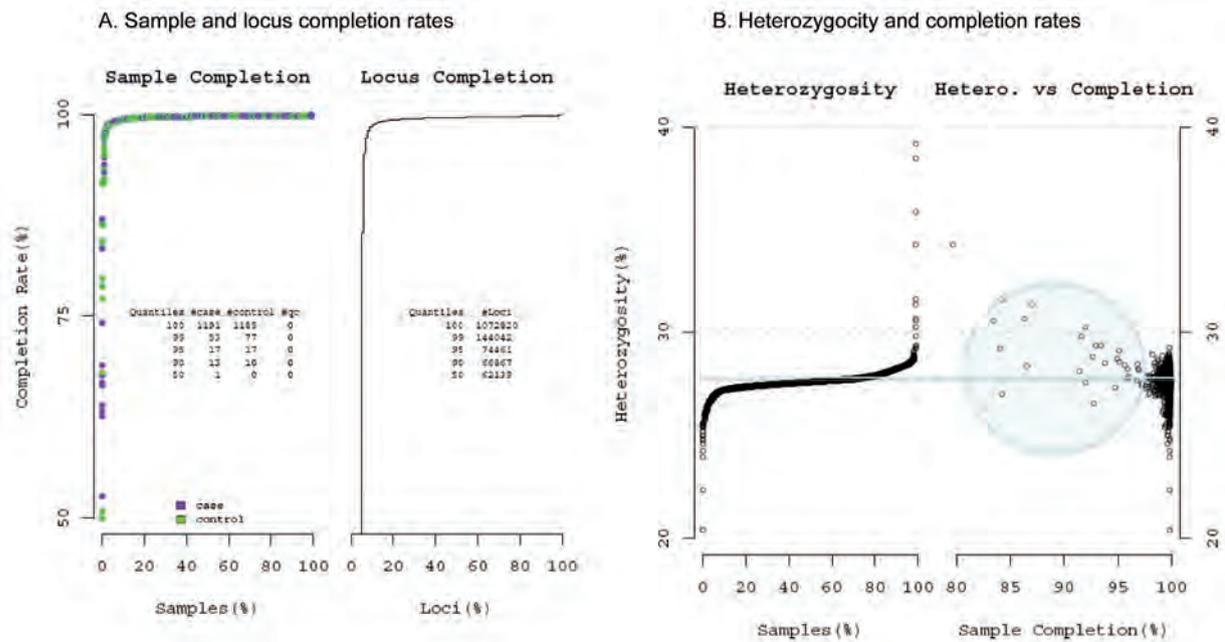


Figure 16.3. Examples of quality control plots using data from 1M Illumina platform



Circle: samples with low completion rates tend to have high heterozygosity values
 Line: average heterozygosity in the population

are p^2 for genotype AA, $2pq$ for Aa and q^2 for aa. Extreme departures from HWP in the control population (departures in cases could be due to associations with disease) can reflect assay problems. Other reasons are also possible, such as non-random mating, selection, population admixture, and a genetically non-homogeneous control population. Therefore, a very careful evaluation of quality control measures should be performed when significant departures from HWP (e.g. using an exact test (32), Pearson's χ^2 test comparing observed and expected genotype frequencies) are observed for a specific SNP assay. Evaluation of HWP for all genotyped SNPs (i.e. comparison of expected and observed number of SNPs with significant HWP departures) can be helpful in determining if the observed departure reflects a problem with the controls, such as a problem during control selection,

or the source population not being in HWE due to non-random mating. If there is no evidence of genotyping errors or control selection problems, the likely explanation for the observed departure is chance. In that case, methods of analyses for associations between the genotype and disease that assume HWP can be helpful in evaluating the impact of a chance departure on estimates of effect, such as the odds ratio (33).

In summary, a list of quality control checks before risk analysis of genotype data can include:

- Verifying duplicate samples and identifying unexpected duplicates using DNA profiling data
- Examining genotype clustering in scatter plots
- Identifying discrepancies between self-reported and genetically determined gender
- Completion rates by sample—excluding data from DNA samples with low completion rates

• Genotype call rates—excluding data from assays with low call rates

• Examining sample heterozygosity and excluding outlier samples

• Genotype concordance among verified duplicate samples (excluding assays with low genotype concordance)

• Genotype concordance among samples not from the same sample (excluding assays with unexpectedly high genotype concordance)

• Testing for deviations of Hardy-Weinberg proportions

Discovery of genetic susceptibility loci

Described here are statistical analyses of SNP data derived from candidate genes or regions, as well as genome-wide approaches in a case-control study (population, hospital-based or nested in a prospective cohort). Methods of

analyses for prospective cohort data (e.g. Cox proportional-hazards regression analyses or analyses of quantitative traits) will not be addressed. The analyses described below can be implemented using widely available statistical packages, such as the commercial packages Stata (<http://www.stata.com/>), SAS (<http://www.sas.com/>) or the software R Project of Statistical Computing (<http://www.r-project.org/>). R is being used more frequently in analyses of genetic data, as many novel statistical methods are written and freely distributed as R add-on packages, providing a very flexible computing and graphical toolset.

Association between individual SNPs and disease risk: Genotype-based analyses

In genotype-based analyses, each individual SNP is evaluated in relation to disease risk by comparing the genotype distribution for cases and controls. The odds ratio (OR) approximates the ratio of disease incidence in exposed (or susceptible) and unexposed (or non-susceptible) individuals, and is often used as a measure of association in case-control studies, as it does not require estimates of the actual incidence rates (34). Table 16.1 illustrates a 2x3 table often used to display the number of cases and controls with the three possible genotypes in the population under study. This data can be used to estimate genotype ORs for subjects carrying the heterozygous and uncommon homozygous genotypes relative to subjects with the common homozygous genotype. Genotype-based ORs can be estimated using logistic regression models (34) with disease status as the outcome and the SNP as the explanatory variable coded as either indicator or dummy

for each genotype. Although these analyses yield unbiased and efficient estimates of the OR, the estimate of the intercept parameter is biased due to the retrospective nature of the case-control design (35). Data from studies with cases individually matched to controls by variables such as age, hospital or region should be analysed by conditional logistic regression models to ensure unbiased and efficient estimates of the OR (34). However, when data on genotype or exposure information is missing for either the case or the control in a matched pair, information from both subjects is lost resulting in decreased efficiency. Therefore, individually matched studies are often analysed as unmatched studies, adjusting for categories of the matching factors using indicator variables. This can result in incomplete adjustment for the matched design, but the impact is generally minimal and compensated for by the gain in efficiency.

Below is the form of a logistic model for genotype (G) variables (additional variables can be added

to adjust for potential confounders) and disease (D) outcome:

$$\Pr(D|G) = \frac{\exp(\beta_0 + \beta_{Aa} Aa + \beta_{aa} aa)}{1 + \exp(\beta_0 + \beta_{Aa} Aa + \beta_{aa} aa)}$$

or

$$\text{logit}(\Pr(D|G)) = \beta_0 + \beta_{Aa} Aa + \beta_{aa} aa$$

where Aa, aa are 0,1 indicator variables for each genotype (AA is the reference).

The genotype-specific OR and 95% confidence intervals (CI) can be estimated from the logistic regression coefficients and its standard error (SE) as:

$$\text{OR}(Aa) = \exp(\beta_{Aa});$$

$$95\% \text{ CI} = \exp(\beta_{Aa} \pm 1.96 \text{ SE}(\beta_{Aa}));$$

$$\text{OR}(aa) = \exp(\beta_{aa});$$

$$95\% \text{ CI} = \exp(\beta_{aa} \pm 1.96 \text{ SE}(\beta_{aa})).$$

The null hypothesis (H_0) for a test of SNP-disease association (sometimes called co-dominant test) can be written as:

$$H_0: \text{OR}(Aa) = 1.0 \text{ and } \text{OR}(aa) = 1.0 \text{ with 2 degrees of freedom (df)}$$

Hypothesis testing can be carried out by conventional score-test, Wald test or likelihood ratio test.

Table 16.1. Genotype frequencies and odds ratio (OR) estimates for genotype-based analyses in a case-control Study

Genotype	Cases	Controls
AA	m	n
Aa	o	p
aa	q	r

m, n, o, p, q, r are cell counts for number of subjects

Genotype-specific estimates

$$\text{OR}(Aa) \text{ vs. } AA = \frac{n * o}{m * p}$$

$$\text{OR}(aa) \text{ vs. } AA = \frac{n * q}{m * r}$$

Dominant model estimate

$$\text{OR}(Aa/aa) \text{ vs. } AA = \frac{n * (o + q)}{m * (p + r)}$$

Variance for genotype-specific OR

$$\text{Var}(\text{LogOR}(Aa)) = \frac{1}{m} + \frac{1}{n} + \frac{1}{o} + \frac{1}{p}$$

$$\text{Var}(\text{LogOR}(aa)) = \frac{1}{m} + \frac{1}{n} + \frac{1}{q} + \frac{1}{r}$$

Recessive model Estimate

$$\text{OR}(aa) \text{ vs. } AA/Aa = \frac{(n + p) * q}{(m + o) * r}$$

The variance of the OR estimates for each genotype is inversely related to each of the cell counts (Table 16.1), which gives an intuitive sense of why the larger the cell counts are, the smaller the variance and the tighter the confidence intervals. This also shows that estimation of genotype-specific ORs can be unreliable for uncommon SNPs with only a few subjects carrying the homozygous variant genotype (aa).

Genotype ORs can also be estimated under the assumption of specific models of genetic inheritance, such as the log-additive (or multiplicative) model, which assumes a log linear trend for genotypes with an increasing number of variant alleles; the recessive model, which assumes the same risk for Aa and aa carriers; and the dominant model, which assumes the same risk for AA and Aa carriers. When using these approaches, keep in mind that these models of inheritance were originally developed for simple Mendelian diseases with near-complete penetrance, and thus might be over simplistic for complex diseases that are influenced by variants in multiple loci. Estimates of ORs and tests for genetic associations under different models for genetic risk can be obtained using logistic regression models with disease status as the outcome. The three possible genotypes for a given SNP are often coded as 0 for AA, 1 for Aa, and 2 for aa. ORs for Aa and aa (relative to AA) can be estimated by including two dummy variables for Aa and aa in a logistic regression model of the form:

$$\text{Logit}(P(D|G)) = \beta_0 + \beta_G G$$

Log-additive trend model: G coded as 0 for AA, 1 for Aa, 2 for aa

$$\exp(\beta_G) = \text{per allele OR}$$

Dominant model: G coded as 0 for AA and 1 for (Aa+aa)

$$\exp(\beta_G) = \text{OR}(Aa/aa) \text{ vs AA}$$

Recessive model: G coded as 0 for

(AA+Aa) and 1 for aa

$$\exp(\beta_G) = \text{OR}(aa) \text{ vs AA/Aa}$$

The H_0 for a test of association under the models above can be written as:

$$H_0: \text{OR} = \exp(\beta_G) = 1.0, 1 \text{ df.}$$

The power to detect disease susceptibility loci is maximized when the assumed genetic model of inheritance is the true model. Thus, since the true mode of inheritance is often unknown, one might choose to test several models (with the caveat mentioned above that simple modes of inheritance might not hold for complex diseases). However, reporting the most significant finding after testing different models will result in an inflated type 1 error or an underestimate of the precision of the confidence intervals. Therefore, it is important to use appropriate statistical tests, such as permutation testing (36), to account for testing of multiple models. Multiple testing might result in no increases or even decreases in power, compared to testing only one pre-defined model with good performance under different alternative models, such as the 2 df genotype-based test or the trend test under the log-additive model (37). The advantage of the 1 df trend test, assuming a log-additive model, is that it uses one less df; it is generally more powerful than the 2 df genotype-based test, when the genetic effect is additive or log-additive; and it has good power to detect dominant effects. However, the trend test has poor power to detect recessive effects. Although the 2 df test for genotype effects has substantially better power to detect recessive effects, the actual power is often low. It should be noted that, even if the underlying model for the disease loci were recessive, the association with disease with a marker SNP (correlated with the causal SNP) would tend to look log-additive due to misclassification of subjects with

respect to the true genotypes for the disease allele. Finally, a 2-sided test for trend is not affected by the sign of the LD between the minor allele of the marker SNPs and the causal SNP, whereas this can be affected for the other tests (38). No matter what model is used for testing for associations, genotype-specific estimates of the OR are often presented (unless the homozygous carriers are very uncommon), since they do not make any assumptions about the underlying mode of inheritance, and provide more information of the possible underlying models.

Association between individual SNPs and disease risk: Allele-based analyses

In genotype-based analyses, the unit of observation is the subject or genotype, whereas in allele-based analyses, the unit of observation is the allele. Since each subject has two alleles at any autosomal locus, the total number of observations in allele-based analyses is twice the number of subjects. Table 16.2 illustrates a 2x2 table for the allele frequencies for cases and controls, and the allele-based ORs and variance. Allele-based tests for association with disease assume independent distribution of alleles in the population, or HWP, for both cases and controls (39). HWP in controls implies HWP in cases only if there is no association with the disease (i.e. under the H_0), or if the alleles have multiplicative (log-additive) effects on disease risk. Therefore, if the control population is in HWP, the allele-based tests with variance estimates under H_0 are valid, but estimation of confidence intervals will require the additional assumption that the alleles have multiplicative effects. It should be noted that under HWP, the

allele-based test is asymptotically equivalent to the trend test in genotype-based analyses (39). The interpretation of the allele-based ORs is less intuitive than the genotype-based ORs, as individuals always carry a combination of two alleles, and thus might not have a useful risk interpretation. Because of the more restricted interpretation and set of assumptions of the allele-based analyses, genotype-based analyses are often preferable (39).

Association between other types of variants and disease risk

The previous section described analyses of SNP variants, which have two alleles for any autosomal loci (males are hemizygotes, i.e. they have only one allele for SNPs in X-linked genes). Analyses of other types of genetic variation, for instance multiallele variants in variable tandem repeats (VTR) or copy number variations (CNV), follow similar principles and will not be discussed here. Because of the increased number of categories, model-based analyses of multiallelic loci or haplotypes can be very helpful in reducing the number of parameters to be estimated (40).

Haplotype analyses

Haplotype analyses exploit the LD, or correlation among genetic markers that are physically close, to improve the statistical efficiency and interpretability of studies of genetic associations with disease risk (41). These analyses can be aimed at comprehensively scanning a candidate region for disease susceptibility loci, or used to detect associations with markers that act in cis (i.e. when two or more variants affect disease only if they are on the same chromosome). A methodological challenge in

Table 16.2. Allele frequencies and odds ratio (OR) estimates for allele-based analysis in a case-control Study

Allele	Cases	Controls
A	2m+o	2n+p
a	2o+m	2p+m

Allele-based Estimates

$$\text{OR}(a) \text{ vs. } A = \frac{(2n+p) * (2o+m)}{(2m+o) * (2p+m)}$$

Variance for allele-specific OR

$$\text{Var}(\text{LogOR}(a)) = \frac{1}{2m+o} + \frac{1}{2n+p} + \frac{1}{2o+m} + \frac{1}{2p+m}$$

studies of unrelated individuals is the estimation of phased haplotypes using unphased genotype data. The estimation of haplotype frequencies among cases and controls is done iteratively using methods such as the expectation-maximization (EM) algorithm, which can be implemented using software packages such as PROC HAPLOTYPE in the SAS Genetics Package, SNPHAP (<http://www.gene.cimr.cam.ac.uk/clayton/software/snphap.txt>) and tagSNP (<http://www-hsc.usc.edu/~stram/tagSNPs.html>). There is a wide range of statistical methods to analyse haplotype associations with disease using regression models that allow for the adjustment of potential confounders (41,42). Single-imputation or “plug-in” methods model estimates of individual haplotypes as if they were observed, whereas marginal regression methods take into account phase ambiguity in the estimation of measures of association between haplotypes and disease risk (41,43). The main advantage of plug-in methods is that they are computationally simple; they can be implemented using standard statistical software,

which uses estimates of posterior haplotype probabilities. More advanced methods use EM type algorithms for simultaneous estimation of haplotype frequencies and haplotype-disease odds ratio parameters. These methods produce more accurate variance estimates and confidence intervals, since they properly account for the fact that haplotype phases are not directly observed. These methods require specialized software, such as the R function haplo.glm in the haplo.stats package, Chaplin (<http://www.genetics.emory.edu/labs/epstein/software/chaplin/>) or HAPSTAT (<http://www.bios.unc.edu/~lin/hapstat/>).

Analysis of GWAS data

GWAS generate very large genotype data sets, often including billions of genotypes per study (e.g. 1 million SNPs in 4000 subjects) that require the development of tools to accommodate the demands for data storage, management, quality control and risk analyses. These tools are likely to expand and improve to meet the needs of

increasingly large data sets. Two free and available tools that can meet the needs of data management, quality control, population stratification and association analyses of GWAS data are PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink>) (44) and the Genotype Library and Utilities (GLU) package (<http://code.google.com/p/glu-genetics>).

Analyses of GWAS data usually include: quality control analysis of genotype data, as described above; analysis of population structure and decisions on the need and method for adjustment for population stratification; definition of the analytical data set after exclusions of data from samples or loci based on quality control analyses; and analysis of the association between genotypes and disease risk.

Analysis of population structure.

Epidemiological studies often collect information on self-reported ethnicity and geographical location from cases and controls. Self-reported ethnicity, race and geographical location are surrogates for a complex mixture of unmeasured factors, which reflect variation in genetic background, culture, language, religion and health-related behaviour (45). Because these unmeasured factors could introduce confounding bias when related to disease and exposures of interest, cases are usually matched to controls by race, ethnicity and geographical location to facilitate adjustment during the analyses. However, population substructure information (i.e. heterogeneous or admixed populations) not captured by self-reported ethnicity could lead to population stratification or confounding bias due to differences in allele frequency and disease risk across subpopulations (46–48). Population structure analyses use multilocus genetic data to assign individuals to populations of origin.

This determines if there is population substructure not accounted for by variables measured in epidemiological studies. Population structure in GWAS can be analysed using a Bayesian clustering approach (49). This method uses information on linkage between a set of SNPs and Hardy–Weinberg disequilibrium to decompose a group of individuals (e.g. cases and controls) into genetically similar populations or clusters. Reference subjects of fixed populations (e.g. Asian, European and African from HapMap) can be used to guide the clustering process, to estimate the degree of admixture of each study sample. The SNPs for these analyses are selected from SNPs in the scan with high completion rates and low residual LD (e.g. $r^2 < 0.1$ – 0.01 for pairs of SNPs less than 200Kb apart). As an outcome of these analyses, each individual is assigned an admixture coefficient reflecting the estimated degree of membership with each population. The degree of membership in a structure analysis, assuming the use of the three HapMap populations as fixed reference populations, can be plotted in an equilateral triangle, also called an admixture plot. Membership estimates for each of the three populations are represented by the distance to each of the three corners of the triangle. The software STRUCTURE can be used to carry out these analyses and can be downloaded from <http://pritch.bsd.uchicago.edu/software.html>.

Genetically-determined race using admixture analyses can be compared to self-reported race and ethnicity to identify, and possibly exclude, outliers from subsequent risk analyses. Figure 16.4A shows an example of an admixture plot generated by STRUCTURE from a population self-identified as

Caucasians. Red dots represent cases and green dots represent controls.

According to this plot, most subjects are estimated to be of more than 85% European descent (i.e. they are clustered in the European corner) with no evidence for substructure. The few subjects that are estimated to be of less than 85% European descent can be excluded from risk analyses to reduce population heterogeneity. However, if strong evidence for population structure were to be found, this method cannot be easily used to adjust for population structure in the risk analyses. EIGENSTRAT is a software that has been proposed to detect and adjust for population stratification in GWAS (50). This method uses principal component analyses (PCA) to reduce high-dimensional genotyping data to lower dimensions that can be used in the analyses. These analyses produce a set of continuous variables, called principal components (PC), that capture the maximum of the genetic variation across individuals in a data set (Figure 16.4B). Each PC is defined as the top eigenvector of a covariance matrix between samples, thus the name EIGENSTRAT. Inclusion of related individuals in PCA analyses can create problems because of the high genetic correlation between relatives. Epidemiological studies of unrelated individuals can occasionally unknowingly enrol family members. Therefore, analyses to determine the degree of relatedness between individuals, such as Pedigree Relationship Statistical Test (PREST) analyses (<http://galton.uchicago.edu/~mcpeek/software/prest/>) (51), should be carried out before the PCA analysis to identify and exclude relatives. Significant PC from these analyses can be used to model ancestry differences between

Figure 16.4A. Example of an triangular admixture plot generated by STRUCTURE for a population self-defined as Caucasians in the USA. The ancestry estimate is represented by the distance to each side of the triangle. Red dots represent cases and green dots represent controls. Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics, copyright (2007).



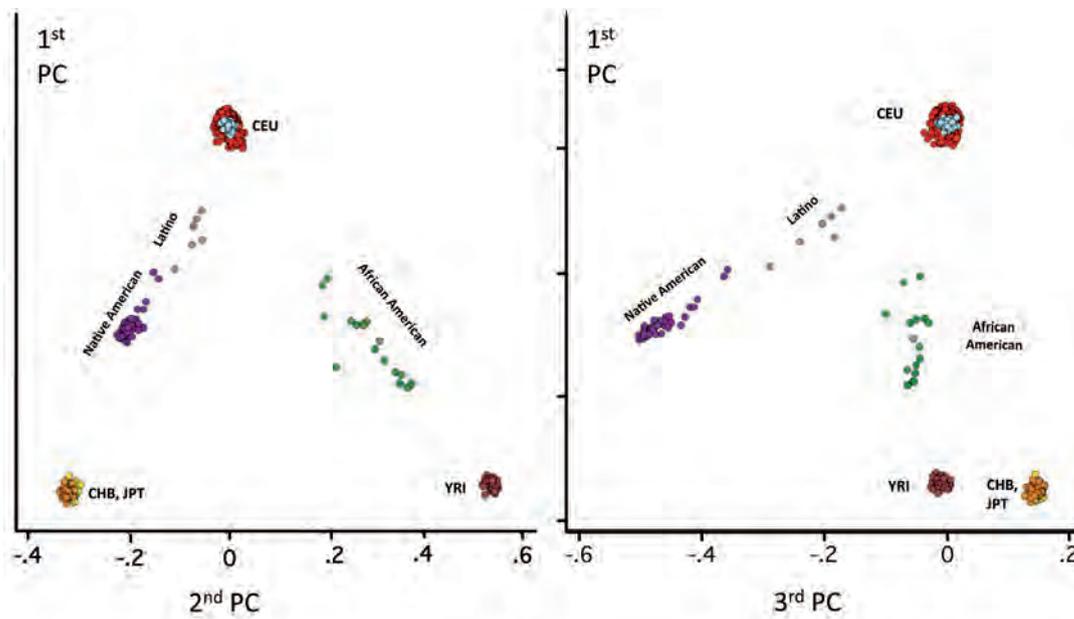
cases and controls and thus adjust for population stratification. These analyses can be performed using the EIGENSOFT package that includes population genetics methods (52) and the EIGENSTRAT stratification

correction method (<http://genepath.med.harvard.edu/~reich/Software.htm>) (50).

SNP imputation methods can increase the power of studies of genetic regions or GWAS by filling

in missing genotype data due to assay failures (e.g. 2% of samples with missing genotypes for a SNP with 98% completion rate), and increasing genetic coverage through imputation of SNPs that have not been genotyped. These methods use information from a reference panel, such as HapMap, to impute untyped SNPs. In addition to increasing power, SNP imputation can sometimes help localize signal for an association in a region, and facilitate the combination of data from studies using different genotyping chips or platforms, including overlapping, but not identical, sets of SNPs. Several methods and software are available to impute SNP data, such as IMPUTE (<http://mathgen.stats.ox.ac.uk/impute/impute.html>) (53), MACH (<http://www.sph.umich.edu/csg/abecasis/MACH/download/>) (54), and BimBam (<http://stephenslab.uchicago.edu/software.html>) (55). In all of these methods, it is important to keep in mind that imputation accuracy will depend

Figure 16.4B. Example of principal component analysis (PCA) to identify principal components (PC) that account for population structure



on many factors: completeness and accuracy to the reference SNP panel; the quality of the data (e.g. large amounts of missing data will decrease accuracy); density of SNPs and LD pattern in the region (i.e. areas with low coverage or low LD are more difficult to impute); the similarity of the LD structure between the reference population (e.g. HapMap) and the population under study (e.g. admixed or unique populations, such as Amish, might be difficult to impute); and allele frequency (i.e. rare SNPs are harder to impute). Thus, findings from analyses of data, including imputed SNPs, should be carefully interpreted taking into account these limitations.

Definition of the analytical data set. Analysis of quality control of the genotype data and population structure can be used to identify and exclude samples and loci from analyses (e.g. samples with low completion rates, samples from subjects with discrepancies between the self-reported race and ethnicity and genetically determined race, or loci with low call rates across samples or with discordant results in duplicated QC samples). Other criteria based on epidemiological data can also be used to define the analytical data set; for instance, exclusion of subjects with missing data in key variables, such as age, or exclusion of rare subtypes of disease, such as rare histological types of cancer, to decrease disease heterogeneity.

Analysis of association between genotypes and disease risk. The primary aim of analysis of GWAS data is to discover markers for genetic susceptibility loci. Initial analyses usually evaluate associations between each individual SNP and disease risk and follow the principles previously described. In multistage designs, an important

consideration in the analysis is the criteria used in the first stage to select SNPs to be genotyped in subsequent stages (10). A subset of the most significant P-values for SNP-disease associations is often used to select SNPs to be carried forward to subsequent stages. Other approaches, such as hierarchical regression models incorporating prior knowledge on the SNP selection procedures, can also be used (56). In general, joint analysis of data from different stages is more powerful than replication analysis (26).

Graphical representation of results can be very helpful for summarizing the large amounts of GWAS data. For instance, quantile-quantile plots (Q-Q plots) for observed P-values for a test of the null hypothesis of no association between each SNP and disease risk against expected P-values under the null hypothesis can be useful. These plots summarize both systematic bias and evidence for association. Most SNPs in a GWAS will not be associated with disease risk; therefore, associated P-values will appear in the diagonal of the Q-Q plot. Small departures at the extreme of the Q-Q plot suggest associations with disease. Large departures from the diagonal can reflect systematic biases leading to increases in false-positive findings (e.g. due to different DNA quality for cases and controls or population stratification). Figure 16.5 shows an example of a Q-Q plot from an analysis of a GWAS of hair colour, before (back dots) and after (red dots) adjusting for population stratification using PC (57). This example shows how adjustment by PC was able to reduce a large deviation from the diagonal that reflected bias due to population stratification. Plots of the $-\log_{10}$ (P-value) for all SNP associations with disease sorted by chromosomal location, can also be

helpful in showing the distribution of P-values and identifying the location of associations with genomic significance.

Analysis of additional outcomes.

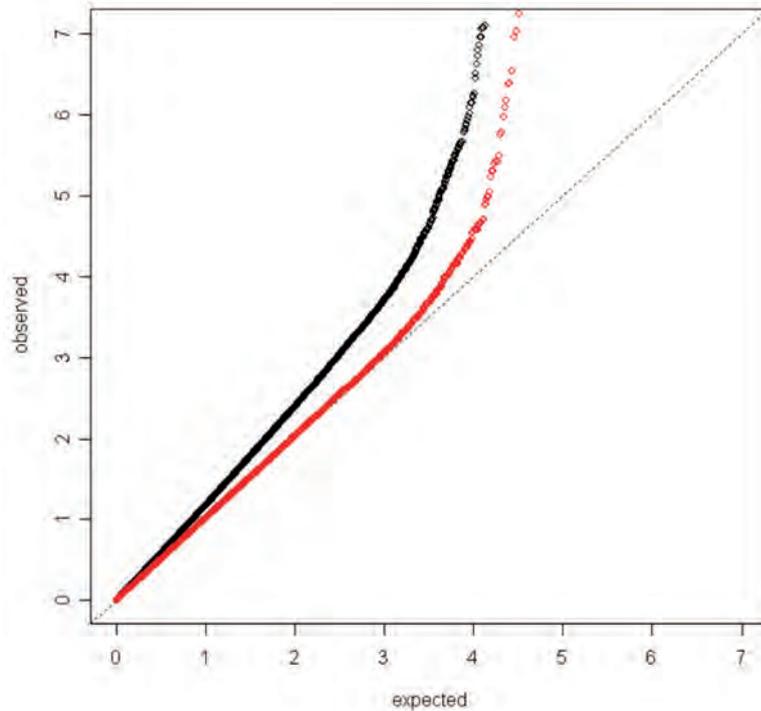
Data from GWAS can also be used in ancillary analyses to evaluate genetic association with secondary outcomes measured in cases and controls (e.g. other diseases or exposures, such as height or smoking habits). The original case-control sampling can affect measures of association with secondary outcomes (58,59); however, bias is only introduced when both the secondary outcome and the genetic loci under study are associated with the risk of the primary disease (58). Because most genetic loci in GWAS will not be associated with the primary disease, naïve analyses, ignoring the sampling design, will be valid for most loci.

Hierarchical-Bayesian methods

Prior information on the expected magnitude of genetic associations (e.g. the OR likely to vary from 1.1–1.5) can be used in hierarchical models to provide more constrained estimates than the conventional, frequentist analytical approaches mentioned above (60,61). Other advantages of this approach are that it can be used to address problems of multiple comparisons and to incorporate biological information from pathway in the analyses.

Hierarchical models can also be used in selecting SNPs to be followed-up in multistage GWAS (60,61). These models can increase the power to detect susceptibility loci by incorporating known information about the SNPs into the selection, rather than just relying on measure of association in the data set. Wider use of hierarchical models has been limited by the unfamiliarity of epidemiologists with software

Figure 16.5. Example of a Quartile-Quartile (Q-Q) plot of observed quartiles of $\log_{10}(\text{P-values})$ against the expected values under the null hypothesis. Black and red dots represent P values from analyses not adjusted and adjusted, respectively, for population stratification using principal components (57).



packages to fit these models. However, their use may increase now that SAS codes are available for analysing epidemiologic data with hierarchical models (60,61).

Evaluation of interactions

Complex diseases are likely to be caused by the interplay of multiple environmental exposures and genetic susceptibilities, hence the importance of evaluating G-G and G-E interactions (62). Specifically, evaluation of interactions can:

- Facilitate the identification of underlying risk factors for disease (e.g. improve power to detect a risk factor that varies according to the levels of another factor by stratifying on the modifier factor)
- Provide insights into the biological mechanisms of disease
- Provide public health benefits,

such as improved risk prediction models and strategies for disease prevention (e.g. benefits of targeting subjects susceptible to specific exposures)

In practice, however, evaluation of interactions can be quite challenging because it requires very large studies with high-quality exposure assessment and availability of biological specimens. Even in well-designed, well-powered epidemiological studies, exploring interactions can be a computationally daunting task, particularly in studies of a very large number of genetic markers, such as GWAS, evaluated in hundreds of thousands of SNPs.

Definition of interaction

In epidemiology, an interaction between two factors is usually

defined as the statistical evaluation of whether the association between one factor (e.g. cigarette smoking) and disease risk varies according to the value of the other factor (e.g. NAT2 genotype). A multiplicative interaction occurs if the association between the two factors is measured in a multiplicative scale by the relative risk (or odds ratio), and an additive interaction occurs if the association is measured in the additive scale by the risk difference (63). A multiplicative or additive interaction can also be described as a departure of the joint effect of the two factors from the expected effect under a multiplicative or additive model, respectively. Table 16.3 shows the definitions of measures of association between two dichotomous factors (an environmental exposure (E) and a genotype (G)) and disease risk, including joint effects and stratum-specific effects, as well as multiplicative and additive interactions (64,65). A set of three ORs characterize the E and G associations with disease: $OR(G|E = 0)$, $OR(E|G = 0)$, and $OR(G,E)$. These ORs can be re-parameterized as the stratum-specific odds ratios and the interaction ORs (i.e. $OR(G|E = 0)$, $OR(G|E = 1)$, and OR_{int} ; or $OR(E|G = 0)$, $OR(E|G = 1)$, and OR_{int}). The relationship between these parameters is shown in Table 16.3.

The biological implications of these two statistical forms of interactions have long been debated in the epidemiologic literature. The main problem in making biological inferences based on epidemiological interactions is that the presence or absence of interaction depends on the scale in which the association with disease is measured. The correspondence between statistical and biological modes for interaction can be defined under simple biological models (62). For instance, under models such as the single-

hit or the sufficient-component-cause, two factors with biologically independent actions on disease result in additive joint effects on the incident rate of the disease (3). However, relationships between biological actions and statistical models in complex diseases with multiple known and unknown causes cannot be easily made, except when the interaction is independent of the scale of measurement of association (66,67). These interactions occur when the effect of one or both factors exists only in the presence of the other, and can be referred to as non-removable interactions. Using notation from Table 16.3, 'non-removable interactions' can be defined as:

- $OR(G|E=0) = 1$ and $OR(E|G=0) = 1$ and $OR(E,G) \neq 1$
- $OR(G|E=0) = 1$ and $OR(E|G=0) \neq 1$ and $OR(E,G) \neq OR(E|G=0)$
- $OR(G|E=0) \neq 1$ and $OR(E|G=0) = 1$ and $OR(E,G) \neq OR(E|G=0)$

The interaction between the NAT2 genotype and smoking status

in bladder cancer risk, where NAT2 slow acetylators are at increased risk of bladder cancer compared to rapid acetylators only among cigarette smokers, is an example of non-removable interactions (68). Crossover, or qualitative interactions, where the effect of one factor is reversed by the presence of the other, is an extreme form of non-removable interactions (69). There are only a few established examples of such interactions in the epidemiologic literature; for instance, the interaction between BMI and menopausal status, where BMI reduces the risk of breast cancer among pre-menopausal women, while it increases the risk among post-menopausal women (70). It is unclear how often G-E or G-G interactions are going to show crossover effects; however, biologically, this extreme type of interaction is generally believed to be rare.

Statistical evaluation of interaction

Interactions between two factors.

Table 16.3 shows different definitions of interactions between two risk factors (either a G-E or G-G interaction) and the null hypotheses that can be tested using data from case-control studies of genetic associations. Logistic regression models, including interaction terms between two or more factors, are commonly used to test multiplicative interactions:

$$\text{logit}(\Pr(D|G, E)) = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G * E$$

$$OR(G|E=0) = \exp(\beta_G)$$

$$OR(E|G=0) = \exp(\beta_E)$$

$$OR(G,E) = \exp(\beta_G) * \exp(\beta_E) * \exp(\beta_{GE})$$

$$OR(E|G=1) = \exp(\beta_E + \beta_{GE})$$

$$OR(G|E=1) = \exp(\beta_G + \beta_{GE})$$

Table 16.3. Odds ratio (OR) estimates for the effects of two binary factors, exposure (E) and genotype (G)

		Genotype		
		0	1	Overall
Exposure	0	1.0 (ref.)	OR(G E=0)	1.0 (ref.)
	1	OR(E G=0)	OR(G,E)	OR(E)
Overall		1.0 (ref.)	OR(G)	

OR definitions	Symbol
Overall or marginal exposure effect	OR(E)
Exposure effect among non-susceptible	OR(E G=0)
Exposure effect among susceptible	OR(E G=1)
Overall or marginal genotype effect	OR(G)
Genotype effect among unexposed	OR(G E=0)
Genotype effect among exposed	OR(G E=1)
Joint genotype and exposure effect	OR(G,E)

Statistical model	No interaction	Interaction parameter
Multiplicative	$OR(G,E) = OR(G E=0) * OR(E G=0)$ $OR(E G=0) = OR(E G=1)$ $OR(G E=0) = OR(G E=1)$	$OR_{int} = OR(G,E) / [OR(G E=0) * OR(E G=0)]$ $= OR(E G=1) / OR(E G=0)$ $= OR(G E=1) / OR(G E=0)$
Additive	$OR(G,E) = OR(G E=0) + OR(E G=0) - 1$	$OR_{int} = OR(G,E) / [OR(G E=0) + OR(E G=0) - 1]$

Re-parameterization to obtain stratum-specific ORs:

$$\text{logit}(\Pr(D|G, E)) = \beta_0 + \beta_{G|E=0} G_0 + \beta_{G|E=1} G_1 + \beta_E E,$$

where G_0 and G_1 are two dummy variables defined as:

$$G_0 = G \text{ if } E = 0; 0 \text{ otherwise}$$

$$G_1 = G \text{ if } E = 1; 0 \text{ otherwise}$$

$$\text{OR}(G|E = 0) = \exp(\beta_{G|E=0}) = \exp(\beta_G)$$

$$\text{OR}(G|E = 1) = \exp(\beta_{G|E=1})$$

Test for multiplicative interaction

$$H_0: \beta_{GE} = 0 \text{ or } \exp(\beta_{GE}) = \text{ORint} = 1$$

In addition to characterizing and testing differences in the relative risk of a factor across levels of another factor, interactions can also be used to increase the power to discover susceptibility loci. This is achieved by accounting for the underlying heterogeneity of the genetic risk due to G-G and G-E interactions (71,72). An omnibus test of the joint null hypothesis of no genetic main effects and interaction (e.g. $H_0: \beta_G = 0$ and $\beta_{GE} = 0$ in model above) can be used for this purpose. Thus, using notation from Table 16.3, one can specify three tests for detecting a genotype effect defined as:

- G-only test: $H_0: \text{OR}(G) = 1$, 1 df
- Subgroup-specific test:
 $H_0: \text{OR}(G|E = 1) = 1$, 1 df
- Omnibus test: $H_0: \text{OR}(G|E = 0) = 1$ and $\text{OR}(G|E = 1) = 1$, 2 df

The power of the omnibus test to detect a genetic effect depends on the precision of both the main effect and the interaction parameter. Therefore, strategies that improve the efficiency of the interaction parameter can increase the power of the omnibus test. For instance, assuming independence between

genetic factors, or between genetic and environmental factors, can lead to important gains in power; however, violation of these assumptions can lead to false-positive findings. Sampling strategies, such as oversampling for uncommon exposures, could interact with genetic markers and also increase the power of the omnibus test. The power advantage of the omnibus test, compared to testing for genetic main effects, is decreased by the presence of error in measuring the interacting exposure (2). The gain in power of the omnibus test with respect to the main effect test is robust to exposure measurement error. For poorly measured exposures, such as diet, there might not be much benefit in accounting for an underlying G-E interaction to detect genetic effects. The disadvantage of the omnibus test derived from the increase in degrees of freedom spent to account for the interaction and the performance of the test is that it can become poor when the degrees of freedom required to model the interaction becomes large. For example, when genetic variation is characterized by tag SNPs within a gene or region, the number of parameters in standard methods required to model interactions with other genes or exposures can become very large. Methods to address this limitation have been proposed (72). Another strategy is to perform multiple omnibus tests for a given genetic factor over a large number of other factors, such as potentially interacting SNPs or exposures. This approach can retain a gain in power, even after adjustment for multiple testing (15,72,73).

The odds ratio interaction parameter can be estimated using only data from cases, if the two interacting factors are independent in the source population of the

cases and the disease is rare in the population (74). This can be easily shown if the ORint is expressed as: $\text{ORint} = \text{OR}_{EG|cases} / \text{OR}_{EG|controls} = 1.0$, where $\text{OR}_{EG|cases}$ is the OR for the association between G and E among cases and $\text{OR}_{EG|controls}$ is the OR for the association between G and E among controls.

If the G-E independence assumption holds (i.e. $\text{OR}_{EG|controls} = 1.0$), then $\text{ORint} = \text{OR}_{EG|cases}$.

An important limitation of this approach is that it does not allow the estimation of other important parameters estimable in case-control data, such as the stratum-specific effects and joint effects of G and E. However, when data from a case-control study is available, assuming independence between interacting factors can be used to increase the power to detect an interaction, without the limitation of the number of parameters that can be estimated (75,76). As in the case-only approach, these methods are subject to severe biases leading to detection of spurious interactions or masking of true interactions if the assumptions are violated. Two-step procedures first test for the G-E independence among the controls, and, based on the acceptance or rejection of the H_0 , a second test uses the case-only (77) or case-control estimator. However, when the G-E association in the controls is modest or the sample size is small, the test in the first step might not have adequate power to reject H_0 . Empirical Bayes methods have been proposed to address the trade-offs between bias and efficiency due to the independence assumption. A comparison of the different approaches mentioned above has been previously described (2,78,79).

Restricting evaluation of interactions only to loci that have previously shown some evidence of an overall association with disease,

independent of the exposures of interest, substantially reduces the possible number of interactions to be evaluated. Of course, the cost of reducing complexity through this approach is that interactions that result in very weak or no overall associations with disease can easily be missed. Data mining techniques that attempt to address this problem are discussed in the next section. Variation in allele frequencies among interacting SNPs can have a strong impact on the power to detect their main effects, which can result in difficulties to replicate findings across populations (73). As larger numbers of epidemiological studies obtain comprehensive genetic data and the methods to evaluate interactions are further developed, the scientific community will be in a better position to characterize complex G-G interactions and evaluate their impact in risk characterization in the population.

Evaluation of high-order interactions

Many of the principles described above for studies on interactions between two factors, such as one G-E or G-G pair, also apply to studies of higher order interactions with three or more risk factors. One of the main difficulties in studying higher order interactions is the complexity of the models to capture relationships between many factors, since the number of possible combinations can be very large. Several data mining methods have been proposed to select models evaluating high order interactions in genetic studies. One method is the traditional stepwise regression approach, which uses statistical significance testing to decide whether higher or lower interaction terms should be kept in the model (80). Other methods

include the Focused Interaction Testing Framework that uses a series of marginal and omnibus tests controlling for false discovery rates to detect susceptible loci (81). Classification and Regression Tree (CART) is a data mining method that is increasingly being used to explore high order G-G and G-E interactions, and can be implemented using the Rpart package in R (82,83). This method uses a recursive-partitioning algorithm that splits a collection of subjects into groups based on the factor that results in the highest discrimination in the disease risk. The procedure starts with all the subjects in the study (root node) and ends with a set of final groups of subjects (nodes) with homogeneous disease risk. The problem of overfitting the data is minimized by cross-validation resulting in “pruning” or “trimming” of the tree. The main limitation of CART is that the resulting model can be very sensitive to peculiarities of the data set being used to generate it, and thus might not be replicated in independent data sets. The output models from CART can be stabilized by bagging, a procedure that combines results from a group or ensemble of trees generated by repeated bootstrap sampling of the data (84). The Random Forest procedure minimizes the correlation between the ensemble of trees by choosing a random subset of factors or exposures for growing the trees in each bootstrap replication (85). A useful feature of these ensemble approaches is that they can generate measures of variable importance of the contribution of each factor on risk, and these measures can be used as an omnibus test statistic capturing both the main effect of a factor and the interactions with other factors. P-values associated with the measures of variable importance can be generated using permutation-

based resampling methods. The randomForest package available in R implements this procedure.

The main feature of logic regression (86), compared to logistic regression models and CART, is that it allows combinations of exposures using “and” and “or” operations rather than only “and.” For instance, in a study evaluating the interaction between SNPs in three loci, a logic regression permits models to have similar risk of disease for subjects with the variant allele in locus 1 and variant alleles in either locus 2 or 3. This specifying “or” operator allows the flexibility of specifying biologically plausible models in which one variant resulting in disruption of a protein product only requires a variant in a class of genetic loci to determine the risk of a disease. In this model, the risk of carrying multiple variants in this class of loci is no higher than just carrying one variant. The optimal logic-tree is determined by cross-validation as in CART. Ensembles of logic trees can be generated by a Markov Chain Monte Carlo method that defines measure of variable importance (87). Logic regressions can be implemented using the LogicReg package in R.

The multifactorial dimension reduction (MDR) non-parametric method has also been proposed to evaluate high order G-G and G-E interactions (88). In contrast to tree-based methods that hierarchically build complex models, MDR reduces the dimensionality of multilocus genotype data by creating binary variables defining high-risk and low-risk groups. This method then evaluates the ability of the derived binary exposure variables to predict disease risk using cross-validation and permutation testing. The parsimony of this method is appealing; however, its performance depends on how well the simple

dichotomization of high-risk and low-risk captures the underlying joint effects of multiple susceptibility loci (88,89). Information software to perform MDR analyses (90) can be found at <http://chgr.mc.vanderbilt.edu/ritchie/lab/method.php?method=mdr>, and an open-source version can be downloaded from <http://www.epistasis.org/software.html>.

The advantage of data mining methods is the flexibility to explore complex, high-order interactions without parametric constraints. However, this can also be a limitation since information on natural or highly plausible constraints is lost, which can result in decreased power and selection of implausible models of interaction. For instance, in studies of G-G interactions it might be reasonable to assume some sort of monotonic trend with increasing number of variant alleles on disease risk. In the case of SNP data, this would mean that the risk of carrying two variant alleles in a given locus is larger than carrying only one variant allele, irrespective of the genotype status of other loci. In logistic regression models, or other parametric models, this constrain is imposed by assuming additive or multiplicative (log-additive) effects of the variant on disease risk. Recent discoveries from GWAS studies provide support for additive or multiplicative effects of genetic markers on disease risk, although these studies might have been underpowered to detect other models, such as recessive mode of inheritance. When evaluating G-E interactions, on the other hand, it might also be reasonable to assume some sort of dose–response relationship between a continuous exposure, such as smoking dose, BMI, or dietary intake of vegetables, and disease risk. This limitation can be addressed by the FlexTree

method, which allows imposing parametric constraints in binary tree-based regression models (91). An R-package to implement this method can be requested at <http://www-stat.stanford.edu/~olshen/flexTree/>.

In summary, data mining methods are promising tools for exploring higher-order G-G and G-E interactions. Their ability to identify reproducible interactions, however, has not yet been demonstrated. Different methods have complementary strengths, and thus the best analytical strategy might be to use a combination of methods and follow-up findings in independent data sets for replication.

Analyses of complex pathways or networks

Candidate genes are often selected from among genes involved in biochemical pathways that are known or thought to be related to the risk factors (e.g. carcinogen metabolizing genes in lung cancer, and other smoking-related cancers). However, the information on how the different genes act in the biological pathway is typically ignored in conventional analyses of the data. As information on the biochemical pathways and networks increases, thanks to the use of profiling or “omics” technologies, such as metabolomics, proteomics and transcriptomics, the interest in incorporating biochemical information in pathway/network analyses of epidemiological studies will grow. Hierarchical-Bayesian methods (92) have been proposed to integrate pathway information into the analyses, although the quantification and integration of biologic information from different sources can be very challenging and potentially limit the usefulness of these approaches. The need for

methodologies for pathway analyses of complex data from molecular epidemiology studies is increasing, and novel methodologies to meet these requirements will likely be developed in the near future.

Concluding remarks

In the coming years, important advances in the understanding of the genetic contribution to complex diseases are likely to be made, facilitated by further advances in genotyping and sequencing technology. The initial discovery of markers of susceptibility in epidemiological studies is just the beginning of new areas of research. Others include:

- Identification of causal genetic variants through fine mapping and functional laboratory studies;
- Evaluation of differences in genotype frequencies and associations with disease in ethnic groups;
- Evaluation of complex interactions and joint effects of multiple susceptibility loci;
- Evaluation of G-E interactions that might facilitate the discovery and characterization of environmental risk factors for disease;
- Evaluation of heterogeneity of genetic associations by disease subtypes;
- Evaluation of the impact of susceptibility loci on individual risk prediction, and identification of population groups with low and high risk of disease; and
- Evaluation of associations between susceptibility loci with additional outcomes, such as disease recurrence, survival and response to therapy.

Therefore, this promising field of research is likely to lead to better understanding of disease etiology, enhancements in risk prediction at the individual and population levels,

and improvements in treatment of disease.

Acknowledgements

Materials from the 2008 Core Genotyping Course on Genetic Analysis (National Cancer Institute, Rockville, MD), and a genetic

analysis course at the 2008 Eastern North American Regional Meeting of the International Biometrics Society, were used for some sections in this chapter. We thank the course instructors for their contributions: Stephen Chanock, Meredith Yeager, Laufry Amundadottir, Sonja Berndt, Nianqing Xiao, Nick Orr, Belynda

Hicks and Amy Hutchinson. We also thank Mark Sherman, Jonine Figueroa and Kelly Bolton for their comments on the chapter.

References

1. García-Closas M, Vermeulen R, Sherman ME *et al.* Application of biomarkers in cancer epidemiology. In: Schottenfeld D, Fraumeni JF Jr, editors. *Cancer epidemiology and prevention*. 3rd ed. New York (NY): Oxford University Press; 2006. p. 70–88.
2. Chatterjee N, Mukherjee B. Statistical approaches to studies of gene-gene and gene-environment interactions. In: Rebbeck TR, Ambrosone CB, Shields PG, editors. *Molecular epidemiology: applications in cancer and other human diseases*. New York (NY): Informa Healthcare; 2008. p. 145–168.
3. Rothman KJ, Greenland S. *Modern epidemiology*. Philadelphia (PA): Lippincott-Raven; 1998.
4. García-Closas M, Lan Q, Rothman N. Design considerations in molecular epidemiology. In: Rebbeck TR, Ambrosone CB, Shields PG, editors. *Molecular epidemiology: applications in cancer and other human diseases*. New York (NY): Informa Healthcare; 2008. p. 1–18.
5. Sham PC, Curtis D (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet*, 59:323–336. doi:10.1111/j.1469-1809.1995.tb00751.x PMID:7486838
6. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992). Selection of controls in case-control studies. I. Principles. *Am J Epidemiol*, 135:1019–1028. PMID:1595688
7. Wacholder S (1991). Practical considerations in choosing between the case-cohort and nested case-control designs. *Epidemiology*, 2:155–158. doi:10.1097/00001648-199103000-00013 PMID:1932316
8. Bergen AW, Haque KA, Qi Y *et al.* (2005). Comparison of yield and genotyping performance of multiple displacement amplification and OmniPlex whole genome amplified DNA generated from multiple DNA sources. *Hum Mutat*, 26:262–270. doi:10.1002/humu.20213 PMID:16086324
9. Antoniou AC, Easton DF (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol*, 25:190–202. doi:10.1002/gepi.10261 PMID:14557987
10. Kraft P, Cox D. Study designs for genome-wide association studies. In: Rao DC, Gu CC, editors. *Genetic dissection of complex traits*. 2nd ed. (Advances in genetics; vol 60). San Diego (CA): Academic Press; 2008. p. 465–504.
11. Johnson GC, Esposito L, Barratt BJ *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29:233–237. doi:10.1038/ng1001-233 PMID:11586306
12. Gabriel SB, Schaffner SF, Nguyen H *et al.* (2002). The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229. doi:10.1126/science.1069424 PMID:12029063
13. Frazer KA, Ballinger DG, Cox DR *et al.*; International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861. doi:10.1038/nature06258 PMID:17943122
14. Carlson CS, Eberle MA, Rieder MJ *et al.* (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74:106–120. doi:10.1086/381000 PMID:14681826
15. Chapman J, Clayton D (2007). One degree of freedom for dominance in indirect association studies. *Genet Epidemiol*, 31:261–271. doi:10.1002/gepi.20207 PMID:17266117
16. Stram DO, Haiman CA, Hirschhorn JN *et al.* (2003). Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered*, 55:27–36. doi:10.1159/000071807 PMID:12890923
17. García-Closas M, Rothman N, Lubin J (1999). Misclassification in case-control studies of gene-environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev*, 8:1043–1050. PMID:10613335
18. García-Closas M, Lubin JH (1999). Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol*, 149:689–692. PMID:10206617
19. Pritchard JK, Przeworski M (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69:1–14. doi:10.1086/321275 PMID:11410837
20. Park JH, Wacholder S, Gail MH *et al.* (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet*, 42:570–575. doi:10.1038/ng.610 PMID:20562874
21. Cardon LR, Bell JI (2001). Association study designs for complex diseases. *Nat Rev Genet*, 2:91–99. doi:10.1038/35052543 PMID:11253062

22. Gail MH, Pfeiffer RM, Wheeler W, Pee D (2008). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics*, 9:201–215. doi:10.1093/biostatistics/kxm032 PMID:17873152
23. Wacholder S, Chanock SJ, Garcia-Closas M *et al.* (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96:434–442. doi:10.1093/jnci/djh075 PMID:15026468
24. Colhoun HM, McKeigue PM, Davey Smith G (2003). Problems of reporting genetic associations with complex outcomes. *Lancet*, 361:865–872. doi:10.1016/S0140-6736(03)12715-8 PMID:12642066
25. Newton-Cheh C, Hirschhorn JN (2005). Genetic association studies of complex traits: design and analysis issues. *Mutat Res*, 573:54–69. PMID:15829237
26. 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. doi:10.1038/ng1706 PMID:16415888
27. Kraft P, Hunter D (2005). Integrating epidemiology and genetic association: the challenge of gene-environment interaction. *Philos Trans R Soc Lond B Biol Sci*, 360:1609–1616. doi:10.1098/rstb.2005.1692 PMID:16096111
28. Ioannidis JP, Bernstein J, Boffetta P *et al.* (2005). A network of investigator networks in human genome epidemiology. *Am J Epidemiol*, 162:302–304. doi:10.1093/aje/kwi201 PMID:16014777
29. Ioannidis JP, Boffetta P, Little J *et al.* (2008). Assessment of cumulative evidence on genetic associations: interim guidelines. *Int J Epidemiol*, 37:120–132. doi:10.1093/ije/dym159 PMID:17898028
30. Rothman N, Garcia-Closas M, Chatterjee N *et al.* (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet*, 42:978–984. doi:10.1038/ng.687 PMID:20972438
31. Chanock SJ. Principles of high-quality genotyping. In: Rebbeck TR, Ambrosone CB, Shields PG, editors. *Molecular epidemiology: applications in cancer and other human diseases*. New York (NY): Informa Healthcare; 2008. p. 63–80.
32. Wigginton JE, Cutler DJ, Abecasis GR (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*, 76:887–893. doi:10.1086/429864 PMID:15789306
33. Chen J, Chatterjee N (2007). Exploiting Hardy-Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. *Hum Hered*, 63:196–204. doi:10.1159/000099996 PMID:17317968
34. Breslow NE, Day NE. Fundamental measure of disease occurrence and association. In: Breslow NE, Day NE, editors. *The analysis of case-control studies*. Lyon: IARC Scientific Publication; 1980. p. 43–81.
35. Prentice RL, Pyke R (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411. doi:10.1093/biomet/66.3.403.
36. Westfall P, Young S. Resampling based multiple testing. New York (NY): John Wiley & Sons, Inc; 1993.
37. Yen YC, Kraft P (2005). Model selection in genetic association studies. *Genet Epidemiol*, 29:289–289.
38. Pfeiffer RM, Gail MH (2003). Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet Epidemiol*, 25:136–148. doi:10.1002/gepi.10245 PMID:12916022
39. Sasieni PD (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53:1253–1261. doi:10.2307/2533494 PMID:9423247
40. Wallenstein S, Hodge SE, Weston A (1998). Logistic regression model for analyzing extended haplotype data. *Genet Epidemiol*, 15:173–181. doi:10.1002/(SICI)1098-2272(1998)15:2<173::AID-GEPI5>3.0.CO;2-7 PMID:9554554
41. Kraft P, Chen J. Haplotype association analysis. In: Rebbeck TR, Ambrosone CB, Shields PG, editors. *Molecular epidemiology: applications in cancer and other diseases*. New York (NY): Informa Healthcare; 2008. p. 205–224.
42. Schaid DJ (2004). Evaluating associations of haplotypes with traits. *Genet Epidemiol*, 27:348–364. doi:10.1002/gepi.20037 PMID:15543638
43. Schaid DJ, Rowland CM, Tines DE *et al.* (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, 70:425–434. doi:10.1086/338688 PMID:11791212
44. Purcell S, Neale B, Todd-Brown K *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81:559–575. doi:10.1086/519795 PMID:17701901
45. Chaturvedi N (2001). Ethnicity as an epidemiological determinant—crudely racist or crucially important? *Int J Epidemiol*, 30:925–927. doi:10.1093/ije/30.5.925 PMID:11689494
46. Wacholder S, Rothman N, Caporaso NE (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst*, 92:1151–1158. doi:10.1093/jnci/92.14.1151 PMID:10904088
47. Thomas DC, Witte JS (2002). Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev*, 11:505–512. PMID:12050090
48. Wacholder S, Rothman N, Caporaso NE (2002). Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev*, 11:513–520. PMID:12050091
49. Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959. PMID:10835412
50. Price AL, Patterson NJ, Plenge RM *et al.* (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38:904–909. doi:10.1038/ng1847 PMID:16862161
51. Sun L, Wilder K, McPeck MS (2002). Enhanced pedigree error detection. *Hum Hered*, 54:99–110. doi:10.1159/000067666 PMID:12566741
52. Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet*, 2:e190. doi:10.1371/journal.pgen.0020190 PMID:17194218
53. Marchini J, Howie B, Myers S *et al.* (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39:906–913. doi:10.1038/ng2088 PMID:17572673
54. Li Y, Mach Abecasis GR (2006). 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet*, 79:2290.
55. Servin B, Stephens M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*, 3:e114. doi:10.1371/journal.pgen.0030114 PMID:17676998
56. Lewinger JP, Conti DV, Baurley JW *et al.* (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*, 31:871–882. doi:10.1002/gepi.20248 PMID:17654612
57. Han J, Kraft P, Nan H *et al.* (2008). A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet*, 4:e1000074. doi:10.1371/journal.pgen.1000074 PMID:18483556
58. Kraft P (2007). Analyses of genome-wide association scans for additional outcomes. *Epidemiology*, 18:838. doi:10.1097/EDE.0b013e318154c7e2 PMID:18049198
59. Richardson DB, Rzehak P, Klensk J, Weiland SK (2007). Analyses of case-control data for additional outcomes. *Epidemiology*, 18:441–445. doi:10.1097/EDE.0b013e318060d25c PMID:17473707
60. Witte JS, Greenland S, Kim LL (1998). Software for hierarchical modeling of epidemiologic data. *Epidemiology*, 9:563–566. doi:10.1097/00001648-199809000-00016 PMID:9730038
61. Chen GK, Witte JS (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet*, 81:397–404. doi:10.1086/519794 PMID:17668389
62. Thompson WD (1991). Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol*, 44:221–232. doi:10.1016/0895-4356(91)90033-6 PMID:1999681

63. Yang QH, Khoury MJ (1997). Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev*, 19:33–43. PMID:9360900
64. Walter SD, Holford TR (1978). Additive, multiplicative, and other models for disease risks. *Am J Epidemiol*, 108:341–346. PMID:727202
65. Clayton D, McKeigue PM (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, 358:1356–1360. doi:10.1016/S0140-6736(01)06418-2 PMID:11684236
66. Ottman R (1996). Gene-environment interaction: definitions and study designs. *Prev Med*, 25:764–770. doi:10.1006/pmed.1996.0117 PMID:8936580
67. Khoury MJ, Beaty TH, Cohen BL, editors. Fundamentals of genetic epidemiology. New York (NY): Oxford University Press; 1993.
68. García-Closas M, Malats N, Silverman D *et al.* (2005). NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, 366:649–659. doi:10.1016/S0140-6736(05)67137-1 PMID:16112301
69. Gail M, Simon R (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41:361–372. doi:10.2307/2530862 PMID:4027319
70. Colditz GA, Baer HJ, Tamimi RM. Breast cancer. In: Schottenfeld D, Fraumeni JF Jr, editors. Cancer epidemiology and prevention. New York (NY): Oxford University Press; 2006. p. 995–1012.
71. Kraft P, Yen YC, Stram DO *et al.* (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*, 63:111–119. doi:10.1159/000099183 PMID:17283440
72. Chatterjee N, Kalaylioglu Z, Moslehi R *et al.* (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet*, 79:1002–1016. doi:10.1086/509704 PMID:17186459
73. Marchini J, Donnelly P, Cardon LR (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37:413–417. doi:10.1038/ng1537 PMID:15793588
74. Piegorsch WW (1994). Statistical models for genetic susceptibility in toxicological and epidemiological investigations. *Environ Health Perspect*, 102 Suppl 1:77–82. PMID:8187729
75. Umbach DM, Weinberg CR (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med*, 16:1731–1743. doi:10.1002/(SICI)1097-0258(19970815)16:15<1731::AID-SIM595>3.0.CO;2-S PMID:9265696
76. Chatterjee N, Carroll R (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92:399–418. doi:10.1093/biomet/92.2.399.
77. García-Closas M, Thompson WD, Robins JM (1998). Differential misclassification and the assessment of gene-environment interactions in case-control studies. *Am J Epidemiol*, 147:426–433. PMID:9525528
78. Mukherjee B, Ahn J, Gruber SB *et al.* (2008). Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol*, 32:615–626. doi:10.1002/gepi.20337 PMID:18473390
79. 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. doi:10.1111/j.1541-0420.2007.00953.x PMID:18162111
80. Millstein J, Conti DV, Gilliland FD, Gauderman WJ (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet*, 78:15–27. doi:10.1086/498850 PMID:16385446
81. Storey J (2002). A direct approach to false discovery rates. *J Royal Stat Soc*, 64:479–498. doi:10.1111/1467-9868.00346.
82. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Pacific Grove (CA): Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
83. Zhang HP, Bonney G (2000). Use of classification trees for association studies. *Genet Epidemiol*, 19:323–332. doi:10.1002/1098-2272(200012)19:4<323::AID-GEPI4>3.0.CO;2-5 PMID:11108642
84. Breiman L (1996). Bagging predictors. *Mach Learn*, 24:123–140. doi:10.1007/BF00058655.
85. Breiman L (2001). Random forests. *Mach Learn*, 45:5–32. doi:10.1023/A:1010933404324.
86. Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic regression. *J Graph Comput Stat*, 12:475–511. doi:10.1198/1061860032238.
87. Kooperberg C, Ruczinski I (2005). Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol*, 28:157–170. doi:10.1002/gepi.20042 PMID:15532037
88. Ritchie MD, Hahn LW, Roodi N *et al.* (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69:138–147. doi:10.1086/321276 PMID:11404819
89. Ritchie MD, Hahn LW, Moore JH (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol*, 24:150–157. doi:10.1002/gepi.10218 PMID:12548676
90. Hahn LW, Ritchie MD, Moore JH (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19:376–382. doi:10.1093/bioinformatics/btf869 PMID:12584123
91. Huang J, Lin A, Narasimhan B *et al.* (2004). Tree-structured supervised learning and the genetics of hypertension. *Proc Natl Acad Sci USA*, 101:10529–10534. doi:10.1073/pnas.0403794101 PMID:15249660
92. Conti DV, Cortessis V, Molitor J, Thomas DC (2003). Bayesian modeling of complex metabolic pathways. *Hum Hered*, 56:83–93. doi:10.1159/000073736 PMID:14614242
93. Hunter DJ, Kraft P, Jacobs KB *et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*, 39:870–874. doi:10.1038/ng2075 PMID:17529973

