8

Statistical Genetic Modeling and Analysis of Complex Traits

Shelley B. Bull
Lunenfeld–Tanenbaum Research Institute and University of Toronto, Toronto, ON

Jinko Graham
Simon Fraser University, Burnaby, BC

Celia M. T. Greenwood
McGill University, Montréal, QC

8.1 Introduction and Overview

The past 15 years have witnessed remarkable developments in the nature and volume of genetic variation data that are available for statistical genetic analysis. Genetic variation generally refers to differences between individuals in the DNA that is inherited from parents. Normally, identical DNA is contained in each of our cells and does not change during a lifetime; it is organized as a string of paired nucleotides for each of 22 chromosomes (autosomes), plus the X and Y sex-chromosomes. A base-pair refers to a pair of nucleotides at a specific position. Simply speaking, a gene can be defined as a set of specific DNA instructions that code an RNA or protein product. These in turn can affect the development of physical features as well as the production of proteins and metabolites that have biological consequences and may eventually play a role in disease causation and physiological variation. The genome refers to all of a person's nucleotides across the chromosomes, and is comprised of 3 billion nucleotides in total, indexed by base-pair position, with roughly 2% within the coding regions of genes, known as exons. As a first step toward discovering and characterizing the role of genes, genetic analysis of DNA variation investigates relationships of specific DNA variants with measurable human traits.

In response to technological advances in measuring DNA variation across the genome of an individual, new ways have arisen to investigate the role of genetic factors in complex traits and this has led to new methods of statistical

modeling and analysis. Our objective in this chapter is to describe some contributions of Canadian researchers to statistical methods in human genetics. We will begin in Section 8.2 by giving a short description of genetic studies that involve either families or groups of unrelated individuals, with an outline of some essentials of the methods. Then in Section 8.3 we will discuss advances in two areas of research, highlighting their applications and impact in disease studies. Section 8.4 will comment on some other current and emerging areas of study. Before doing this, we will first describe some of the ways that information on a person's genome is obtained.

In the investigation of the genetics of complex traits and diseases, in which trait variation and disease risk arise from a combination of multiple genetic and environmental factors, the ultimate scientific objective is to understand the underlying genetic and biological mechanisms. This process often begins by scanning (i.e., screening) the entire genomes of specific individuals to identify a chromosomal region that may harbor a genetic variant that is a risk factor for the disease or explains variation in the trait of interest. The two primary methods of analysis used in such genome-wide studies are genetic linkage and genetic association analysis. Genome-wide studies involve measurement of a large number of genetic markers, each with a known genomic position (locus) on a specific chromosome. At an observed genetic locus, the measured marker can take on different values or variants, known as alleles.

As summarized in Table 8.1, early genome scans used sets of markers known as microsatellites that can take on many different values (referred to as polymorphisms), but these were subsequently replaced with cost-effective arrays of single nucleotide polymorphisms (SNPs) which are markers that take on only two values (e.g., alleles A and a) at a single base-pair position on a chromosome. The design of these arrays exploits information about local dependence between neighboring SNP markers (known as linkage disequilibrium, LD) to choose so-called tagSNPs which serve to represent the genetic information in a small region, saving the effort of measuring every SNP therein. A tagSNP usually has no biological role, but can indirectly detect variation at an unobserved SNP that is directly involved in disease expression. SNP arrays are designed to comprehensively assess the kind of genetic variation across the genome that occurs reasonably often in human populations, both within genes and in regions outside genes. The density of SNPs measured per chromosome has steadily increased with each improvement in array technology, so that standard arrays can measure more than one million SNP markers genome-wide, with estimates of another 2 million by genetic imputation based on LD information available from external population sources through the International HapMap Project. Next generation sequencing (NGS) technology, now emerging, aims to directly measure variation at every base-pair position with sufficient accuracy for near complete characterization of an individual's genome, including variants that occur very rarely in a population.

TABLE 8.1: Summary of technologies for measurement of genetic variation. Microsatellite markers have been widely used in genetic analysis of study designs involving pedigrees, affected relatives, and case-parent trios. SNP arrays now predominate in most study designs, including those with unrelated individuals, particularly case-control studies, and use of next generation sequencing is increasing.

High-Throughput Technology	Type of Variant	Number of Markers/ Variants	Genotyping Accuracy
Microsatellite Markers	Highly polymorphic	Hundreds	High
SNP Arrays	Binary: common	1 Million	Moderate to High
Next Generation Sequencing	Binary: Rare, low frequency and common	3 Billion	Variable

8.2 Essentials of Statistical Genetic Methods

By examining patterns of inheritance in families, genetic linkage analysis aims to detect chromosomal regions containing genes that influence the risk of specific inherited diseases or traits. We need to know the pedigree (relationships among family members), as well as disease or trait information and genetic typing for at least some of the family members. Two genetic loci are said to be linked when the parental alleles transmitted to a child at one locus are not independent of the parental alleles at the other locus. Figure 8.1 illustrates the genetic transmission of chromosomal material to the children of two parents drawn from a randomly mating population (i.e., within a nuclear family).

In Figure 8.1(a), the patterned vertical bars represent a pair of chromosomes for each of six individuals in a population. The horizontal tick marks represent chromosome positions at which a genetic marker or variant can be measured; one allele occurs on each chromosome. As a simple example, suppose that the dark diagonal and light dotted chromosomes each carry the A allele at a marker locus on this chromosome, and the other chromosomes carry the a allele at the same locus. A genotype for an individual is composed of the pair of unordered alleles observed at the marker locus, for example AA, Aa,

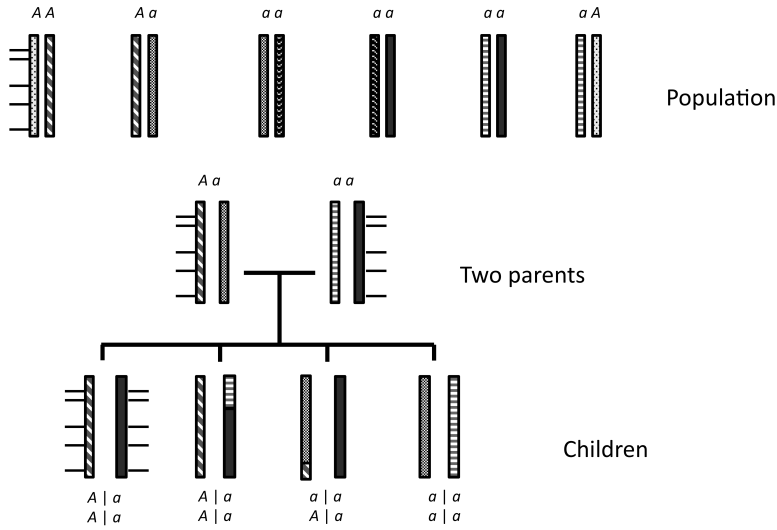

FIGURE 8.1: Illustration of (a) the genetic variability in a population, as well as (b) genetic transmission from parents to offspring in a nuclear family (two parents and their children) and identical-by-descent (IBD) sharing among siblings.

or aa. Genotyping methods provide information about the two alleles at the particular location, but not the parental source (so that Aa is not distinguishable from aA). In the nuclear family depicted in Figure 8.1(b), which consists of parents and their children, the parents drawn from the population produce four children. A chromosome composed of two patterns represents a recombinant chromosome, while a chromosome of one pattern is non-recombinant. A recombinant chromosome can be created by the occurrence of crossovers during gamete formation (either the egg or the sperm) in which chromosomal material is exchanged between the two chromosomes of one parent, although exchange does not always happen. For each gamete, such crossover events occur independently with varying probabilities and locations. Each child then inherits one chromosome from their mother, and likewise, one from their father. When a genetic marker is close to a disease gene, it is more likely that a parental allele at the marker locus is jointly transmitted with a specific allele from the same parent at the disease locus. Informative parental genotyping, here represented by differently-patterned chromosomes, is usually required for precise inference about which parent transmitted a particular allele.

One type of genetic linkage analysis, usually referred to as model-based or parametric linkage analysis, requires explicit model assumptions about the relationship between the unknown gene variants and the disease or trait, and

the pattern of disease inheritance in a family. Another type of linkage analysis, known as model-free analysis, does not require specification of a disease inheritance model; see Xu et al. (2012) for a comprehensive summary. Rather, patterns of genetic similarity among affected relatives at a marker locus are compared to what one would expect at a location distant from a disease gene. Genetic similarity is typically assessed by allele sharing, which is discussed in the following subsection. Genetic linkage tests have been used as a first step in the search for disease susceptibility genes, and a genome-wide scan for linkage may simply identify broad regions that harbor such genes. These regions can then be examined more closely by analysis with alternative methods or by collection and analysis of higher density genetic typing data.

Genetic association analysis also aims to detect chromosomal regions that harbor genes for complex traits, but proceeds by testing the null hypothesis of no association between a SNP genotype and disease status, or trait value. Case-control studies of unrelated individuals that compare the distribution of genotypes in those with and without disease (i.e., "affected" and "unaffected") have predominated. Nevertheless, statistical methods for genetic association analysis involving families have remained of considerable interest. Family-based designs for assessing genetic association in families include the case-parent trio design, in which an affected child and his or her parents are genotyped. Under this design, one examines whether an affected child inherits a parental allele more often than expected under Mendelian inheritance, or in other words, whether allele transmission from the parents in a sample of trios exceeds the expected proportion of 1/2. Family-based designs have been extended to the analysis of disease status and trait values in nuclear families and large pedigrees, combining comparisons between unrelated individuals with comparisons between related individuals from the same family (Mirea et al., 2012).

Recently and with the advent of next generation sequencing (NGS), there is renewed interest in classical genetic linkage analysis methods. In families carrying rare genetic mutations that nearly always produce disease, linkage analysis can efficiently narrow down the genomic regions likely to carry the causal variant. Consequently, there is a substantial increase in the accuracy of inferring which of the genetic alterations identified by NGS is most likely to be a disease-causing mutation.

8.2.1 Modeling Genetic Sharing in Sibships and Relative Pairs

Here, we focus our discussion on genetic linkage analysis for a binary disease state (affected/unaffected), and we assume that sufficient genetic data are available so that it is possible to quite accurately infer whether two related individuals share 0, 1 or 2 copies of a chosen chromosomal region. Such sharing implies that the chromosomal region was inherited from a common ancestor, and this is termed identical-by-descent (IBD) sharing of chromosomal regions.

In the nuclear family illustrated in Figure 8.1, the first and the second child share the dark diagonal chromosome inherited from one parent but share only the lower (solid dark) part of the chromosome inherited from the other parent, so IBD sharing is equal to 1 in the upper part of the chromosome and equal to 2 in the lower part. The second and third child have IBD sharing equal to 0 in the upper part of the chromosome, IBD sharing equal to 1 in the middle part, and IBD sharing equal to 2 at the lower tip. In contrast, the first and the last child have IBD sharing equal to 0 across the entire chromosome. An observation in this family that the first, second, and third child were all affected with disease, whereas the last was not, would be consistent with a disease gene locus in the lower tip of the chromosome.

Mendel's law says that a child is equally likely to inherit one half or the other half of a parent's chromosomal material. In sib pairs generally (i.e., not selected according to their disease status), or in affected sib pairs at markers that are distant from a disease gene locus, we expect IBD sharing proportions of $(1/4, 1/2, 1/4)$ for IBD sharing of $(0, 1, 2)$ copies, respectively. (At a specific location, for example at the top of the chromosomes in Figure 8.1(b), there are four possible genotypes that are equally likely to be inherited by a child. Then for a pair of siblings, there are 16 possible combinations of two genotypes, and among these, there are four that share IBD = 0, eight that share IBD = 1, and four that share IBD = 2). However at a marker close to a disease susceptibility gene, we expect to observe excess IBD in a sample of affected sib pairs, with a higher proportion of IBD = 2 and a lower proportion of IBD = 0.

Conceptually, the goal of genetic linkage analysis is to find chromosomal regions with excess IBD sharing among related individuals. Such a chromosomal region is then likely to contain a gene or genetic variant that alters disease risk. In the absence of any association with disease, the probability distribution of IBD sharing follows from Mendel's laws of inheritance. We denote the probability that a pair of related individuals share k alleles identical by descent as $z_k = \Pr(\text{share } k \text{ copies})$. For a randomly chosen chromosomal region (i.e., a region in which no disease gene is located), the IBD sharing probabilities are $(z_0, z_1, z_2) = (1/4, 1/2, 1/4)$ for siblings, and $(3/4, 1/4, 0)$ for first cousins, for example. When multiple markers within a chromosome region are examined for linkage in a sample of affected relatives, a location with sufficiently large excess IBD sharing is inferred to be close to a potential disease gene locus, while more distant loci exhibit lesser amounts of excess sharing.

Figure 8.2 illustrates the characteristic reduction in average IBD sharing in affected siblings with increasing distance from a linkage peak at which IBD sharing is maximal. Marker locations under the linkage curve of excess sharing constitute a linkage region.

Statistical estimation and testing of hypotheses is often done using likelihood functions (LF). A likelihood function associated with a particular dataset is obtained by considering the probability of the observed data under a hypothesized model; it thus is a function of the parameters that specify the model. Here the z_k are the parameters in the likelihood for a simple allele

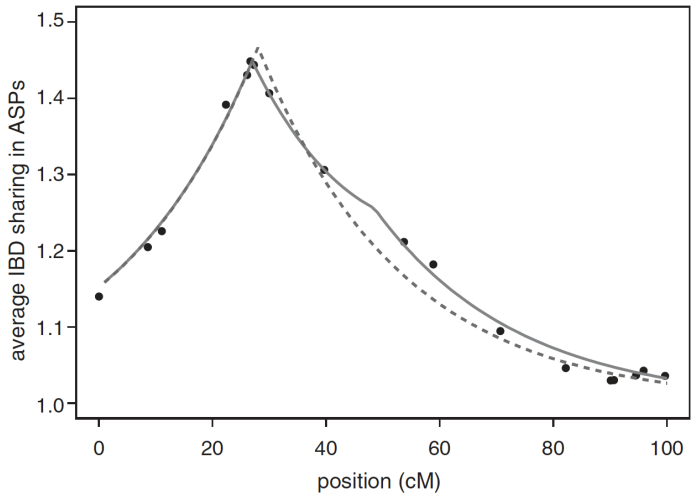

FIGURE 8.2: Patterns of identical-by-descent (IBD) sharing in a region of linkage on chromosome 6 in a sample of affected sib pairs (ASPs) with diabetes (taken from Biernacka et al., 2005). The solid circles indicate average IBD sharing values in the ASPs at each of 18 markers. The two curves show the expected IBD sharing for a one-locus model (dashed line) and for a two-locus model (solid line); cM is a measure of chromosome marker position commonly used in linkage analysis.

sharing linkage model, and they can be estimated and tested using the LF. Formally, one way to test for linkage is by a likelihood ratio (LR) statistic where the likelihood at the estimated IBD proportions $z = (\hat{z}_0, \hat{z}_1, \hat{z}_2)$ is compared to the likelihood under the null hypothesis of no excess sharing. The LR statistic can be maximized with respect to the z_k parameters, and the null hypothesis of no linkage is rejected when the maximized LR statistic is large in comparison to a test criterion. The maximum likelihood estimation is constrained such that the estimated IBD proportions sum to 1. In affected sibling pairs, the power of the LR test to detect linkage can be improved by evaluating the likelihood at a different point z^*, where z^* is constrained to lie within the so-called plausible triangle of z_k values ($z_1 \leq 1/2$ and $z_1 \geq 2z_0$) that are consistent with the underlying genetics of IBD allele sharing (Holmans, 1993; Feng et al., 2005).

8.2.2 Models for Genetic Transmission in Families

Family-based designs for genetic association are also useful in the analysis of disease status in families, and analysis can be conducted in a nuclear family

TABLE 8.2: Formulation of allele transmission from parents to an affected child, for family-based association tests of excess transmission under a case-parent trio design: n_{Aa} is the observed count of Aa parents that transmit A and not a, n_{aA} is the observed count of aA parents that transmit a and not A, n_{AA} is the observed count of homozygous AA parents, and n_{aa} is the observed count of homozygous aa parents.

	Not Transmitted	
Transmitted	A	a
A	n_{AA}	n_{Aa}
a	n_{aA}	n_{aa}

with only one affected child. Returning to Figure 8.1, now we consider that child 1 is affected and with two parents together constitutes a case-parent trio. Suppose that the dark diagonal chromosome carries the A allele at a marker locus on this chromosome, and the other chromosomes carry the a allele at the same locus. When we genotype a marker locus on the chromosome, we find that the first parent transmitted their diagonal chromosome A allele and did not transmit their gray chromosome a allele to the child, whereas the second parent transmitted their solid dark a allele and not their striped a allele. More generally, assuming n trios with $2n$ parents, we can construct a 2×2 table as given in Table 8.2 in which n_{Aa} is the observed count of parents that transmit A and not a, n_{aA} is the observed count that transmit a and not A, and so on. In our simple example, parent 1 would contribute to the n_{Aa} count, but parent 2 is homozygous for a and contributes to the n_{aa} count. If the marker allele is unrelated to being affected with disease then the probability that a parent transmits one allele or the other is $1/2$, and among parents with the Aa genotype, no difference is expected in the proportions of A and a alleles that are transmitted. In contrast, excess transmission of the A allele with n_{Aa} greater than n_{aA}, for example, would provide family-based evidence for association of that allele with disease. As in the allele-sharing analysis, the success of this approach is dependent on the presence of linkage between the marker locus and the disease locus.

In the case of a biallelic genetic variant such as a SNP (e.g., alleles A and a) and a single affected child, a test statistic to detect excess transmission can be constructed from the off-diagonal cell counts in the 2×2 table, where the margins correspond to transmitted and untransmitted alleles for the parents, and each parent contributes one observation to the table (Table 8.2). Among several model formulations developed for this kind of pattern of genetic inheritance, known as transmission disequilibrium, one that lends itself well

to generalization is the conditional on parental genotypes (CPG) likelihood (Schaid and Sommer, 1993). Following the notation of Mirea et al. (2012) for the case-parent trio design, this likelihood models $\Pr(G \mid G_m, G_f, D = 1)$, the probability of the child's genotype G conditional on the child being a case ($D = 1$) and the parental genotypes (G_m for the mother and G_f for the father), in terms of the genotype relative risk parameters

$$\psi_{AA} = \frac{\Pr(D = 1 \mid G = AA, G_m, G_f)}{\Pr(D = 1 \mid G = aa, G_m, G_f)}$$

and

$$\psi_{Aa} = \frac{\Pr(D = 1 \mid G = Aa, G_m, G_f)}{\Pr(D = 1 \mid G = aa, G_m, G_f)}.$$

Assuming a gene-dose model where

$$\psi_{Aa} = \psi \quad \text{and} \quad \psi_{AA} = \psi_{Aa}^2 = \psi^2,$$

the maximum likelihood estimate is $\hat{\psi} = n_{Aa}/n_{aA}$. Families in which both parents are homozygous (e.g., AA or aa) at a marker locus are uninformative. The null hypothesis specifying no excess transmissions, $\mathcal{H}_0 : \psi = 1$, can be formally tested using a LR statistic.

8.2.3 Genetic Association in Unrelated Individuals

The era of genome-wide association studies (GWAS) has been largely driven by the availability of inexpensive SNP-based technology to systematically examine the entire genome. Statistical analysis in GWAS is characterized by the application of conventional methods for hypothesis testing of each one of a very large number of SNPs genotyped using arrays (see Chapter 9 by Craiu and Sun for discussion of approaches to statistical inference in GWAS). The problem of high false positive error rates generated by conducting millions of tests has been addressed by requiring strict significance thresholds in the discovery study and replication in an independent study. Because stringent criteria for genome-wide statistical significance are necessary to reduce the number of false positive associations, very large sample sizes are required, and case-control study designs have predominated; these studies compare the genotypes of unrelated individuals affected with disease (i.e., cases) to those of healthy individuals from the general population or individuals known to be unaffected with the disease (i.e., controls).

As illustrated in Table 8.3, assuming n cases and m controls have been genotyped at a SNP locus, we can construct a 2×3 table to compare their genotype frequencies. We denote the genotype probabilities for cases and controls by (p_2, p_1, p_0) and (q_2, q_1, q_0) respectively, according to the number of copies of the less frequent allele in the genotype (e.g., the number of copies of A in AA, Aa or aA, aa). Assuming a gene-dose model in which the probability

TABLE 8.3: Formulation of genetic association in unrelated individuals under a case-control design: n_2 is the observed count of cases with genotype AA, n_1 is the observed count of cases with genotype Aa, n_0 is the observed count of cases with genotype aa, and m_x is the observed count of controls with x copies of allele A.

	Individual Genotype		
	AA	Aa	aa
Case	n_2	n_1	n_0
Control	m_2	m_1	m_0

of being affected depends on the number of copies of the allele, a genotypic odds ratio parameter, OR, can be defined such that

$$\text{OR} = \frac{p_2/q_2}{p_1/q_1} = \frac{p_1/q_1}{p_0/q_0},$$

and estimated from the observed genotype counts using likelihood methods.

GWAS provide a cost-effective platform to detect associations, and can often identify a narrower chromosome region than linkage analysis. However, association analysis has been largely limited to SNP variants that occur in more that 5% of individuals, and a disease-associated GWAS tagSNP is not usually causal itself but is close to a causal variant. GWAS typically have not provided sufficient refinement at the base-pair level to identify potential disease-causing variants. Next generation sequencing (NGS) and genetic imputation to augment the set of genotyped SNPs can be implemented in GWAS samples to better determine the base-pair location of the genetic variant that is possibly relevant. In some studies, only the regions surrounding the significant tag SNPs are sequenced. In other studies, the entire genome or exome is imputed or sequenced. The ability to comprehensively examine all genetic variation, including that occurring rarely or at low frequency in a population, is the main attraction of NGS for genetic association studies, but in this chapter we leave aside discussion of methods for the analysis of NGS rare variants, which is currently a very active area of research.

8.3 Advances in Statistical Genetic Methods

8.3.1 Linkage with Covariate Data

Allele-sharing models such as the one introduced in Section 8.2.1 have been widely used for the study of genetic linkage in complex traits and diseases. When the etiology of a disease is complex (i.e., when there are many factors coming together to cause disease), heterogeneity can occur because of differences in genetic or environmental factors that cause disease, or because individuals with disease subtypes having different causes appear to be affected with the same disease. In this situation, families may exhibit excess IBD sharing at several marker loci and subsets of the families may be linked to different markers. Likewise, if an environmental exposure as well as a high-risk gene variant need to co-occur in order to increase risk of disease, then unexposed families will not exhibit excess sharing even for a marker close to the disease gene locus. Classifying families according to covariate data (i.e., data on factors that may be disease-related) can often help to reduce this heterogeneity, and improve the sensitivity of linkage analysis. The inherent challenge in considering covariate effects on genetic linkage is that a covariate is defined for an individual, but linkage is defined by examining sharing in pairs or sets of individuals. It is therefore necessary to construct covariates that apply to a pair or a set. For example, in a pair of diseased siblings, it might be of interest to know if they were both diagnosed at a young age, with the idea that genetic factors are more important in early onset than in adult onset disease. The concept of heterogeneity in linkage evidence was instrumental in a study of genetic factors influencing inflammatory bowel disease. Canadian families containing at least two individuals diagnosed with either Crohn's disease (CD) or ulcerative colitis (UC) were recruited and a linkage analysis based on IBD allele sharing in 158 families identified a region of interest on chromosome 5 (Rioux et al., 2000). Exploration of covariate associations in affected relative pairs in an overlapping but slightly larger set of families (167 families and 199 affected sibling pairs) found evidence for differences in allele sharing that depended on diagnostic subtype and age at diagnosis (Bull et al., 2002). The sharing patterns in the families were associated more strongly with CD than with UC, and particularly with families containing at least one individual with a young age at diagnosis.

For simplicity, the discussion here assumes that one chromosomal region is being examined and that the genetic data for an individual at that location are represented by G. If the probability of being diagnosed with disease D depends on a covariate, x, such as age or an environmental exposure, as well as the genotype G, this can be thought of as a modification of the conditional probabilities $\Pr(D|G)$ by the covariate; i.e.,

$$\Pr(D|G, x_1) \neq \Pr(D|G, x_2),$$

where x_1 and x_2 are different covariate values. Given this assumption, it can then be shown that the probability distribution of IBD genetic sharing between a pair of diseased, related individuals also depends on the covariate values of the pair; calculations for these probabilities essentially use Bayes rule. Let x_P represent a covariate that applies to a pair of affected relatives, and $k = 0, 1, 2$ enumerate the IBD states. In one of the first studies in this direction (Greenwood and Bull, 1999a), the dependence of the IBD proportions on covariates in affected sibling pairs was parameterized by a multinomial model,

$$z_k(x_P) = \frac{\exp(\beta_k x_P)}{1 + \exp(\beta_1 x_P) + \exp(\beta_2 x_P)}$$

in which the parameters β_1 and β_2 correspond to differences in IBD sharing that depend on covariates. In the study of CD and UC families noted above, both family-level and pair-level covariates were defined (i.e., ethnic background: Jewish versus non-Jewish; diagnostic subtype: all sibs CD, all sibs UC, or both CD and UC sibs; and age at diagnosis: one sib diagnosed by age 16 years versus both sibs diagnosed after age 16). Although inference about linkage in the absence of covariate effects can be improved by the triangle constraints introduced in Section 8.2.1, their use in linkage models with covariates is more complicated, since the constraints may no longer apply for all sibling pairs, particularly for pairs who have different values for the covariate. The most attractive option may be to implement partial constraints, for example within covariate-defined subgroups, which although not optimal, may nevertheless improve power of the likelihood ratio test in most situations (Greenwood and Bull, 1999a).

Subsequent development of approaches to test for linkage in the presence of covariates beyond the affected sibpair analysis we have described here allow many types of relative pairs to be analyzed simultaneously. For example, because the expected IBD sharing proportions depend on the type of relative pair, Xu et al. (2006) extended the model of Olson (1999) to include relative-pair-level covariates. In this framework, the likelihood contributions of different relative pair types are unified by writing the probabilities in terms of the increased disease risk for IBD $= k$ ($k = 1$ or 2) relative to IBD $= 0$. The resulting covariate-dependent relative risk formulation could then be incorporated into a tree-based algorithm designed to select the important covariates; this method identified clinical covariates altering evidence for linkage at a candidate region in families including cases of bipolar affective disorder (Xu et al., 2006). It is also possible to include individual-level covariates in IBD-sharing-based linkage analysis through judicious weighting of each individual's contribution to the test of linkage (Whittemore and Halpern, 2006). Finally, for linkage studies that examine multiple pairs of relatives in the same family, practical methods to account for the dependence between the pairs provide accurate assessment of statistical significance (Greenwood and Bull, 1999b; Schaid et al., 2007).

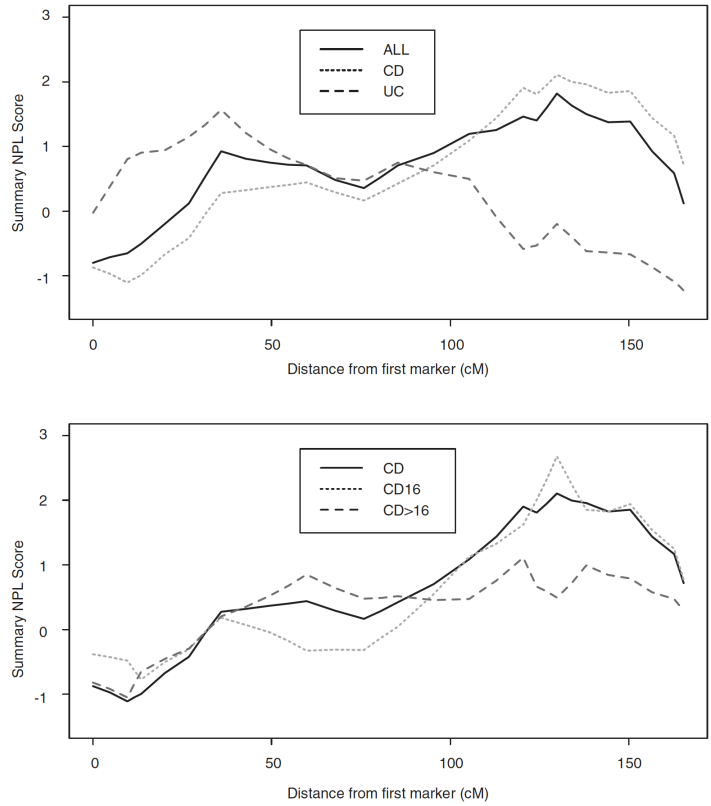

FIGURE 8.3: Summary nonparametric linkage (NPL) scores for family subgroups with inflammatory bowel disease in a region of linkage on chromosome 5 (taken From Mirea et al., 2004). The solid line in the upper figure applies to all families, CD and UC denote diagnostic subgroups, while CD 16 and CD>16 distinguishes subgroups defined by minimum age at onset ≤ 16 years in the family. The summary NPL score is an average of family-specific linkage statistics calculated using 17 markers. cM is a measure of chromosome marker position commonly used in linkage analysis.

Further analysis of the inflammatory bowel disease families at the chromosome 5 locus using family-level tests of covariate-based heterogeneity (Mirea et al., 2004) also demonstrated some evidence for heterogeneity in linkage depending on the pedigree minimum age of onset (Xu et al., 2012). The nonparametric linkage (NPL) statistic used in Figure 8.3 is a summary of excess IBD sharing among all the affected relatives in a family. We see differences between CD and UC subtypes in the family-specific NPL score, and according to early age at diagnosis in the CD subtype. The prominent genomic location

in the latter has shown consistent association with disease risk in subsequent studies (Weizman and Silverberg, 2012), but unfortunately, identifying exactly which DNA changes are responsible for the increased disease risk has proven extremely elusive. Investigation of the genetics of inflammatory bowel disease continues, with a recent large meta-GWAS study of over 75,000 cases and controls, yielding a cumulative total of 163 loci that meet genome-wide significance thresholds (Jostins et al., 2012).

8.3.2 Effect Estimation in Genome-Wide Analysis

Whether conducted by linkage or association analysis, genome-wide investigations are fundamentally discovery studies in that they aim to comprehensively examine all regions of the genome and discover regions, genes, or variants that will be evaluated in subsequent studies. It is important for scientific acceptance that such potential effects also be seen in one or more independent "replication" studies of individuals or families that did not participate in the discovery study. When the same observations are used for both gene discovery and effect estimation, and a genetic effect is estimated only when the test for linkage or association at that locus meets genome-wide statistical significance criteria, the effect estimate is on average larger in magnitude than the true value (Göring et al., 2001). Bias in genetic effect estimates, a phenomenon also known as the winner's curse or the Beavis effect (Xu, 2003), can occur in both genome-wide linkage and genome-wide association studies. The bias introduced in this way is a form of selection bias (Efron, 2011), and the observation of a smaller effect in an independent replication sample is common. This phenomenon is well-illustrated by a case-control GWAS of psoriasis that reported odds ratios for genetic association calculated in independent discovery and replication samples (Nair et al., 2009). All of the 10 most significant SNPs identified in the discovery sample of 2759 individuals had smaller odds ratios in the replication sample of 10,099 individuals, with the percentage reduction ranging from 5% to 20%.

Such bias is particularly critical in replication study design. When an estimate from a discovery study is used to calculate the sample size required for a replication study, a biased effect estimate will produce a sample size estimate that is too small, leaving the investigators with a study under-powered to replicate a true association. If the association is not replicated, it may be assumed the original finding was a false positive association, when in fact it is a true association that was not replicated due to inadequate sample size. In addition, an accurate estimate of the genetic effect is important for estimation of the proportion of heritability explained by significant genetic associations, for meta-analyses from multiple studies, and for interpretation of detected associations.

To address the problem of selection bias, initially in the setting of allele-sharing linkage analysis, Sun and Bull (2005) developed a genome-wide "bootstrap" resampling method in which random samples drawn from the origi-

nal study data are reanalyzed. Within each such bootstrap resample, which consists of a subset of the original study observations, repeating the entire genome-wide analysis imitates a "discovery" study, and captures the effects of the statistical significance criterion, as well as the correlation structure among the genetic loci. A genetic effect estimate for a locus with a test statistic that exceeds the genome-wide significance threshold in the bootstrap resample analysis, will be subject to the "winner's curse." Then to imitate an independent study, the genetic effect at the locus discovered in the bootstrap resample is estimated in the original study observations not included in the bootstrap resample. The difference between these two estimates reflects the winner's curse bias. This process is repeated in each of a large number of bootstrap resamples and the average difference is taken as an estimate of bias. Because the size of the bias depends on the ranking of the loci according to the size of the test statistic (i.e., whether it is the largest or the tenth largest, for example), the bias for the kth ranked locus detected in the original analysis is estimated with the bias estimates for the kth ranked locus in each of the bootstrap resamples. This is called a shrinkage estimator because the bias calculated in this way is used to reduce the genetic effect estimate for each of the loci identified in the original study.

This bootstrap implementation has proven valuable in several scientific settings, including genetic linkage analysis, GWAS and NGS (Wu et al., 2006; Yu et al., 2007; Faye et al., 2011, 2013). Wu et al. (2006) evaluated the bootstrap method for a quantitative trait linkage scan and, in the situation where multiple significant markers are detected, described how to estimate the proportion of genetic variance explained by a marker (a quantity known as locus-specific heritability). In GWAS, it can account for SNP selection according to statistical significance or relative ranking among SNPs, and a software package with an efficient implementation for GWAS is publicly available (`www.utstat.utoronto.ca/sun/Software/BR2/br2-web/br2.html`). Applications in genetic association studies of psoriasis, diabetic glycemia, and time to diabetic nephropathy have provided additional insight into the performance and prospects for replication studies (Al-Kateb et al., 2008; Paterson et al., 2010; Sun et al., 2011; Poirier et al., 2013). Additional aspects of selection bias have been studied in the NGS setting in which regions surrounding highly ranked GWAS SNPs are sequenced by NGS in the same sample used to detect the region (Faye and Bull, 2011). In addition, SNP genotyping is subject to both measurement error and to errors of imputation. Recent work has considered further the consequences of differential sequencing error and imputation error, and applied the bootstrap shrinkage estimator in an analytic re-ranking method that improves identification of potentially causal variants, for example, in locating sequence variants within a chromosome region associated with the risk of prostate cancer (Faye et al., 2013).

8.4 Commentary: Current and Emerging Issues

For complex traits, such as cancer and obesity, epidemiologic evidence implicates environment and lifestyle as major factors. On the other hand, familial clustering of disease suggests that genetics also plays a role. An emerging focus of complex-trait epidemiology is the role of the environment (E) together with genes (G), often referred to as $G \times E$, or gene-environment interaction. In early genome-wide association studies (GWAS), environmental and lifestyle factors were rarely considered, owing to challenges with the feasibility, cost and validity of covariate measurement. However, GWAS have explained very little of the variation in traits that is due to genetic differences (Manolio et al., 2009), and so investigators are taking a closer look at $G \times E$ (Ober and Vercelli, 2011), with some success. For instance, among smokers, colorectal cancer is strongly associated with exposure to carcinogens in well-done red meat, but only in genetically susceptible individuals (Le Marchand et al., 2001). As another example, obesity is more highly associated with fat intake in carriers than in non-carriers of the risk allele in the PPAR-γ gene (e.g., Garaulet et al., 2011). In fact, the concept of $G \times E$ was instrumental in a recent longitudinal study of body-mass index (Abarin et al., 2012). The study followed 1096 children from birth to 14 years of age and analyzed the variation in their BMI growth trajectories. Among carriers of the risk allele in the fat mass and obesity gene, *FTO*, the increase in BMI was attenuated by longer duration of exclusive breastfeeding. These $G \times E$ findings have public health relevance for preventing obesity in genetically susceptible individuals.

Despite the successes described in this chapter, statistical challenges remain. For example, in cancer and other rare diseases, the case-control study is a practical design, but the ability to detect $G \times E$ is low, even with large sample sizes (Greenland, 1983). When the study population is ethnically homogeneous and E is not genetically influenced, it is reasonable to assume that G and E are statistically independent; i.e., a person's genotype is unrelated to their environmental covariates. In these cases, power may be enhanced by enforcing G–E independence (e.g., Umbach and Weinberg, 1997; Chatterjee and Carroll, 2005; Shin et al., 2007), as implemented in the R software package `LUCA` (Shin et al., 2007), available from the Comprehensive R Archive Network (CRAN; `cran.r-project.org`). Outside genetics, these methods can also be applied to exploit the independence of randomized treatments and covariates in nested case-control studies (Breslow et al., 2013). One drawback is that even slight departures from the assumed form of G–E dependence can inflate the chance of false-positive results in tests of $G \times E$ (e.g., Shin et al., 2007). Statistical independence of G and E within families, rather than in the population, is a weaker assumption that can increase the power to detect $G \times E$ in case-parent study designs. These designs are frequently used for rare diseases of early onset, such as childhood cancers (e.g., Infante-Rivard et al., 2007) in

which prenatal factors such as parental occupational exposures may be relevant. Shin (2012) developed a data-driven approach that incorporates G–E independence within families to visualize and test $G \times E$ in case-parent trios. The approach is implemented in the R software package `trioGxE` available on CRAN.

Measurement or misclassification error can affect the interpretation of results for $G \times E$ in ways that require more research in terms of identification and correction (Greenland, 1980). Development of approaches to appropriately deal with the misclassification of G (e.g., Weinberg et al., 2011; Shin et al., 2012) and with E (Thomas, 2010) is an important area of ongoing research. Recent commentaries (Thomas, 2012; Thomas et al., 2012) suggest that methods for the joint analysis of genes and environment are likely to take on increasing prominence, in part due to the development of molecular technologies and biomarkers that can broadly assess environmental exposures.

Acknowledgments

We would like to mention here, with thanks, our trainees, colleagues, and collaborators: David Andrews, Joanna Biernacka, Andy Boright, Laurent Briollais, Mary Corey, Gerarda Darlington, Linnea Duke, Laura Faye, Claire Infante-Rivard, Sophia Lee, Juan Pablo Lewinger, Brad McNeney, Lucia Mirea, Kenneth Morgan, Andrew Paterson, Thomas Schulze, Jean Shin, Mark Silverberg, John Spinelli, Lei Sun, Dave Tritchler, Longyang Wu, and Wei Xu. In addition, we wish to acknowledge the following sources of funding: Mathematics of Information Technology and Complex Systems, Natural Sciences and Engineering Research Council, Canadian Institutes of Health Research.

About the Authors

Shelley B. Bull is a senior investigator in the Lunenfeld–Tanenbaum Research Institute of Mount Sinai Hospital, and a professor of biostatistics in the Dalla Lana School of Public Health, University of Toronto, where she co-directs the CIHR Strategic Training Program in Advanced Genetic Epidemiology and Statistical Genetics. Her undergraduate degree in mathematics and master's degree in statistics are from the University of Waterloo, and her PhD in epidemiology and biostatistics is from the University of Western Ontario. She is the recipient of the Anthony Miller Award for Excellence in Research in Public Health and the International Genetic Epidemiology Leadership Award. She was program chair of the Statistical Society of Canada Annual Meeting in 2011.

Jinko Graham is an associate professor of statistics and Actuarial Science at Simon Fraser University. Her research is directed toward the development of statistical methods for inference from genomic data, with a focus on genetic association studies and applications of the coalescent. Her undergraduate degree in mathematics and master's degree in statistics are from the University of British Columbia, and her PhD in biostatistics is from the University of Washington. From 2005 to 2012, she served on the CIHR Institute of Genetics Priority and Planning Committee for Population Genetics, Genetic Epidemiology and Complex Diseases.

Celia M. T. Greenwood is a senior scientist at the Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, and an associate professor at McGill University, with affiliations to three departments: Oncology, Epidemiology, Biostatistics and Occupational Health, and Human Genetics. Her research interests include numerous aspects of methodology for the analysis of human genetic and genomic data. She has degrees from McGill, the University of Waterloo, and the University of Toronto. After several years at the Hospital for Sick Children and the University of Toronto, she was appointed Weekend to End Cancer Career Scientist at the Lady Davis Institute in 2010.

Bibliography

Abarin, T., Wu, Y. Y., Warrington, N., Lye, S., Pennell, C., and Briollais, L. (2012). The impact of breastfeeding on FTO-related BMI growth trajectories: An application to the Raine pregnancy cohort study. *International Journal of Epidemiology*, 41:1650–1660.

Al-Kateb, H., Boright, A. P., Mirea, L., Xie, X., Sutradhar, R., Mowjoodi, A., Bharaj, B., Liu, M., Bucksa, J. M., Arends, V. L., Steffes, M. W., Cleary, P. A., Sun, W., Lachin, J. M., Thorner, P. S., Ho, M., McKnight, A. J., Maxwell, A. P., Savage, D. A., Kidd, K. K., Kidd, J. R., Speed, W. C., Orchard, T. J., Miller, R. G., Sun, L., Bull, S. B., Paterson, A. D., and Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group (2008). Multiple SOD1/SFRS15 variants are associated with the development and progression of diabetic nephropathy: The DCCT/EDIC Genetics study. *Diabetes*, 57:218–228.

Biernacka, J. M., Sun, L., and Bull, S. B. (2005). Simultaneous localization of two linked disease susceptibility genes. *Genetic Epidemiology*, 28:33–47.

Breslow, N. E., Amorim, G., Pettinger, M. B., and Rossouw, J. (2013). Using the whole cohort in the analysis of case-control data. *Statistics in Biosciences*, 5:232–249.

Bull, S. B., Greenwood, C. M. T., Mirea, L., and Morgan, K. (2002). Regression models for allele sharing: Analysis of accumulating data in affected sib pair studies. *Statistics in Medicine*, 21:431–444.

Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92:399–418.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106:1602–1614.

Faye, L. L. and Bull, S. B. (2011). Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: Accuracy of genetic effect estimates. *BMC Proceedings*, 5 Suppl 9:S64.

Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B., and Sun, L. (2013). Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genetics*, 9(8):e1003609. doi:10.1371/journal.pgen.1003609.

Faye, L. L., Sun, L., Dimitromanolakis, A., and Bull, S. B. (2011). A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. *Statistics in Medicine*, 30:1898–1912.

Feng, Z. Z., Chen, J., and Thompson, M. E. (2005). The universal validity of the possible triangle constraint for affected sib pairs. *The Canadian Journal of Statistics*, 33:297–310.

Garaulet, M., Smith, C., Hernández-González, T., Lee, Y., and Ordovás, J. (2011). PPAR-γ Pro12Ala interacts with fat intake for obesity and weight loss in a behavioural treatment based on the Mediterranean diet. *Molecular Nutrition & Food Research*, 55:1771–1779.

Göring, H. H., Terwilliger, J. D., and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *American Journal of Human Genetics*, 69:1357–1369.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 112:564–569.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: A review and a study of power. *Statistics in Medicine*, 2:243–251.

Greenwood, C. M. T. and Bull, S. B. (1999a). Analysis of affected sib pairs, with covariates–with and without constraints. *American Journal of Human Genetics*, 64:871–885.

Greenwood, C. M. T. and Bull, S. B. (1999b). Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests, under the assumption of no linkage. *American Journal of Human Genetics*, 64:1248–1252.

Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *American Journal of Human Genetics*, 52:362–374.

Infante-Rivard, C., Vermunt, J., and Weinberg, C. R. (2007). Excess transmission of the NAD(P)H:Quinone Oxidoreductase 1 (NQO1) C609T polymorphism in families of children with acute lymphoblastic leukemia. *American Journal of Epidemiology*, 165:1248–1254.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491:119–124.

Le Marchand, L., Hankin, J. H., Wilkens, L. R., Pierce, L. M., Franke, A., Kolonel, L. N., Seifried, A., Custer, L. J., Chang, W., Lum-Jones, A., and Donlon, T. (2001). Combined effects of well-done red meat, smoking, and rapid n-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiology Biomarkers & Prevention*, 10:1259–1266.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Mirea, L., Briollais, L., and Bull, S. B. (2004). Tests for covariate-associated heterogeneity in IBD allele sharing of affected relatives. *Genetic Epidemiology*, 26:44–60.

Mirea, L., Infante-Rivard, C., Sun, L., and Bull, S. B. (2012). Strategies for genetic association analyses combining unrelated case-control individuals and family trios. *American Journal of Epidemiology*, 176:70–79.

Nair, R. P., Duffin, K. C., Helms, C., Ding, J., Stuart, P. E., Goldgar, D., Gudjonsson, J. E., Li, Y., Tejasvi, T., Feng, B.-J., Ruether, A., Schreiber, S., Weichenthal, M., Gladman, D., Rahman, P., Schrodi, S. J., Prahalad, S., Guthery, S. L., Fischer, J., Liao, W., Kwok, P.-Y., Menter, A., Lathrop, G. M., Wise, C. A., Begovich, A. B., Voorhees, J. J., Elder, J. T., Krueger, G. G., Bowcock, A. M., Abecasis, G. R., and Collaborative Association Study of Psoriasis (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics*, 41:199–204.

Ober, C. and Vercelli, D. (2011). Gene-environment interactions in human disease: Nuisance or opportunity? *Trends in Genetics*, 27:107–115.

Olson, J. M. (1999). A general conditional-logistic model for affected-relative-pair linkage studies. *American Journal of Human Genetics*, 65:1760–1769.

Paterson, A. D., Waggott, D., Boright, A. P., Hosseini, S. M., Shen, E., Sylvestre, M.-P., Wong, I., Bharaj, B., Cleary, P. A., Lachin, J. M., MAGIC (Meta-Analyses of Glucose and Insulin-related traits Consortium), Below, J., Nicolae, D., Cox, N. J., Canty, A. J., Sun, L., Bull, S. B., and Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Research Group (2010). A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both HbA1C and glucose. *Diabetes*, 59:539–549.

Poirier, J., Faye, L. L., Dimitromanolakis, A., Paterson, A. D., Sun, L., and Bull, S. B. (2013). A general procedure to address the winner's curse in genetic association studies: Bias-reduction in analysis of time-to-event traits. *Submitted for publication*.

Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., McLeod, R. S., Griffiths, A. M., Green, T., Brettin, T. S., Stone, V., Bull, S. B., Bitton, A., Williams, C. N., Greenberg, G. R., Cohen, Z., Lander, E. S., Hudson, T. J., and Siminovitch, K. A. (2000). Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. *American Journal of Human Genetics*, 66:1863–1870.

Schaid, D. J., Sinnwell, J. P., and Thibodeau, S. N. (2007). Testing genetic linkage with relative pairs and covariates by quasi-likelihood score statistics. *Human Heredity*, 64:220–233.

Schaid, D. J. and Sommer, S. S. (1993). Genotype relative risks: Methods for design and analysis of candidate-gene association studies. *American Journal of Human Genetics*, 53:1114–1126.

Shin, J.-H. (2012). *Inferring gene-environment interaction from case-parent trio data: Evaluation of and adjustment for spurious G×E and development of a data-smoothing method to uncover true G×E*. Doctoral Dissertation, Simon Fraser University, Burnaby, BC. Available at `http://summit.sfu.ca/item/12281\#310`.

Shin, J.-H., Infante-Rivard, C., Graham, J., and McNeney, B. (2012). Adjusting for spurious gene-by-environment interaction using case-parent triads. *Statistical Applications in Genetics and Molecular Biology*, 11:1–21.

Shin, J.-H., McNeney, B., and Graham, J. (2007). Case-control inference of interaction between genetic and nongenetic risk factors under assumptions on their distribution. *Statistical Applications in Genetics and Molecular Biology*, 6:Article 13.

Sun, L. and Bull, S. B. (2005). Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28:352–367.

Sun, L., Dimitromanolakis, A., Faye, L. L., Paterson, A. D., Waggott, D., DCCT/EDIC Research Group, and Bull, S. B. (2011). BR-squared: A practical solution to the winner's curse in genome-wide scans. *Human Genetics*, 129:545–552.

Thomas, D. C. (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annual Review of Public Health*, 31:21–36.

Thomas, D. C. (2012). Genetic epidemiology with a capital E: Where will we be in another 10 years? *Genetic Epidemiology*, 36:179–182.

Thomas, D. C., Lewinger, J. P., Murcray, C. E., and Gauderman, W. J. (2012). Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome. *American Journal of Epidemiology*, 175:203–7; discussion 208.

Umbach, D. M. and Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 16:1731–1743.

Weinberg, C. R., Shi, M., and Umbach, D. M. (2011). A sibling-augmented case-only approach for assessing multiplicative gene-environment interactions. *American Journal of Epidemiology*, 174:1183–1189.

Weizman, A. V. and Silverberg, M. S. (2012). Have genomic discoveries in inflammatory bowel disease translated into clinical progress? *Current Gastroenterology Reports*, 14:139–145.

Whittemore, A. S. and Halpern, J. (2006). Nonparametric linkage analysis using person-specific covariates. *Genetic Epidemiology*, 30:369–379.

Wu, L. Y., Sun, L., and Bull, S. B. (2006). Locus-specific heritability estimation via the bootstrap in linkage scans for quantitative trait loci. *Human Heredity*, 62:84–96.

Xu, S. (2003). Theoretical basis of the Beavis effect. *Genetics*, 165:2259–2268.

Xu, W., Bull, S. B., Mirea, L., and Greenwood, C. M. T. (2012). Model-free linkage analysis of a binary trait. In *Statistical Human Genetics*, volume 850 of *Methods in Molecular Biology*, pp. 317–345. Humana Press, New York.

Xu, W., Schulze, T. G., DePaulo, J., Bull, S. B., McMahon, F., and Greenwood, C. M. T. (2006). A tree-based model for allele-sharing-based linkage analysis in human complex diseases. *Genetic Epidemiology*, 30:155–169.

Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N., and Wacholder, S. (2007). Flexible design for following up positive findings. *American Journal of Human Genetics*, 81:540–551.