January 2, 19

Genetic Epidemiology

Draft chapter for *Modern Epidemiology* (4th Edition)

John S. Witte and Duncan C. Thomas

INTRODUCTION

Genetic epidemiology marries the fields of human genetics and epidemiology. It covers detection of the genetic origin of phenotypic variability in humans (Vogel, 2000), inquiry into the genetic components that contribute to the development or progression of diseases or traits, and evaluation of environmental or other risk factors that may modify the effects of genetic factors.

The first reference to the term "genetic epidemiology" was in a 1954 textbook (J. V. Neel & Schull, 1954), but it was not until the founding of the International Genetic Epidemiology Society in 1983 that the term came to be widely used. In an editorial accompanying the inaugural issue of the society's official journal, Rao (1984) wrote:

> Genetic epidemiology differs from epidemiology by its explicit consideration of genetic factors and family resemblance; it differs from population genetics by its focus on disease; it also differs from medical genetics by its emphasis on population aspects.

Rao's is just one interpretation of the term "genetic epidemiology". In a second editorial, Neel (1984) commented on genetic epidemiology investigating the multifactorial nature of many chronic diseases and highlighted a concern about environmental mutagens. Later, Thomas (D Thomas, 2000) summarized three distinguishing characteristics of the field: a focus on population-based research; joint consideration of genes and the environment; and incorporation of biological processes into its conceptual models. Only by large-scale appraisal of genetics, environment, and their interaction can a full understanding of the etiology of complex traits be achieved.

Genetic epidemiology is closely related to molecular epidemiology (Chapter xx of this volume), but the former has historically focused on germline and somatic genetic factors, while the latter has focused on cellular and molecular biomarkers (D C Thomas, 2004). This distinction has become increasingly blurred as both fields have transitioned into an era of *integrative genomics*. Investigators are using a range of measurements (e.g., epigenetics, transcription, metabolism, protein synthesis) to understand complex pathways involving both genetic and environmental exposures. Emerging high-volume and high-density omics technologies, as discussed below, have made these evaluations possible.

Genetic epidemiology research relied heavily on the development of novel statistical methods and has been critical to our understanding of human diseases. Most early successes involved the discovery of single genes that caused rare Mendelian disorders such as cystic fibrosis (Kerem et al., 1989) and Huntington's disease (MacDonald et al., 1991). Even these seemingly simple etiologies were soon recognized as more complex, however, with the discovery of multiple mutations within causal genes (Cordovado et al., 2012) and modifier genes affecting severity (F. A. Wright et al., 2011). It has become clear that many diseases and traits are complex—such as cancer or height—in that they are attributable to a large number of genetic variants, each with a small effect on risk (Fisher, 1918). Deciphering the genetic basis of such polygenic (Frank Dudbridge, 2016; Purcell et al., 2009)—or even *omnigenic* (Boyle, Li, & Pritchard, 2017)—traits may require very large populations and novel analytic approaches (Thun, Hoover, & Hunter, 2012).

The incorporation of environmental factors into studies of genetics has been slower-moving. While some complex diseases have long been recognized as having both environmental and familial components, our understanding of their interaction is only in its infancy. Early insights in breast cancer, for example, include the potential interaction of cigarette smoking with the *NAT2* gene (Ambrosone et al., 1996) and of ionizing radiation with the *ATM* gene (J. L.

Bernstein et al., 2010). However, a full understanding of how genes and environment affect breast cancer remains elusive (Fejerman et al., 2015). While genetic epidemiology has detected many important associations between genetic variants and a broad range of diseases and traits, much more remains to be understood.

METHODOLOGIES AND CHALLENGES IN GENETIC EPIDEMIOLOGY

Fundamental Principles

An understanding of key terminology is fundamental to comprehension of genetic epidemiology studies. It is also critical to appreciate the basic principles of the transmission of genetic information to later generations and across populations. In addressing these fundamental concepts below, we emphasize binary disease traits (e.g., cancer) for simplicity, but the general principles are equally applicable to continuous quantitative traits (e.g., blood pressure) with appropriate model specification.

General Terminology

The *genome* can be defined as the complete collection of an individual's genetic material. It consists of *chromosomes*, which are long strands of deoxyribonucleic acid (DNA) present in most cells. The human genome is *diploid*, whereby chromosomes are paired. Every human cell contains 22 *autosomal* pairs (which are *homologous*) and one pair of sex chromosomes (X and Y). During *meiosis*, a diploid chromosome set is reduced to a *haploid* chromosome set of a germ cell, the *gamete*.

A *gene* is a piece of a chromosome that codes for a function. Regions that encompass genes (~1-2% of the genome) and non-genes (~98-99% of the genome) can have important effects on disease. A *locus* is a position along the chromosome that denotes the location of a *genetic marker*. For simplicity's sake, we refer to any piece of DNA as a marker, whether it is in a genic or non-genic region. Regions in the genome exhibiting variation across individuals are said to contain *variants*, the different versions of which are called *alleles*. The smallest variant is a *single nucleotide polymorphism* (SNP), or if rare, called a *single nucleotide variant* (SNV). Each pair of autosomal chromosomes contains the same markers, with possibly different alleles, at the same locations. The pair of an individual's alleles at a marker is called a g*enotype*. If the alleles are identical, the individual is called *homozygous* at the marker; otherwise the individual is called *heterozygous*. When considering several loci simultaneously, the multi-locus alleles inherited from the same parent constitute a *haplotype*.

Genetic Transmission

Genetic information is transmitted from each generation to the next according to the fundamental laws of Mendelian inheritance. One copy of each pair of chromosomes is randomly transmitted from each parent's respective gametes (sperm or ova), which upon fertilization form a complete genotype.

Segments that have identical DNA sequences in two or more individuals are called *identical by state (IBS)*. Such segments are considered to be *identical by descent (IBD)* if the individuals inherited them from a common ancestor without recombination. *IBS* segments can alternatively result from the same mutation arising multiple times in the past. Any two individuals can share 0, 1, or 2 alleles IBD with probabilities that depend upon their familial relationship. For example, full siblings have IBD probabilities equal to ¼, ½, and ¼ for sharing 0, 1, or 2 alleles, respectively. In contrast, parent-offspring pairs always share exactly 1 allele

IBD. The selection of which allele is transmitted from a parent to his or her offspring is purely random; this phenomenon is termed Mendel's *law of segregation*.

The relationship between a genotype and disease is termed *penetrance* and defined by Mendel's *law of dominance*. If only one of the two alleles determines the phenotype, that allele is called *dominant* and the other *recessive*. If both alleles have an effect on disease risk, inheritance is said to be *co-dominant*, and when the effect of two different alleles on disease is midway between the effect of homozygotes for either of the two alleles, inheritance is said to be *additive*.

Markers on separate chromosomes or located far apart on the same chromosome segregate independently; this is Mendel's *law of independent assortment*. However, markers located nearby on the same chromosome may be transmitted together, a process known as *genetic linkage*, with a probability that depends upon the distance between them (the *recombination fraction,* θ). This phenomenon underlies linkage analysis, a method for mapping disease genes by studying the co-segregation of the disease with markers in families.

Linkage Disequilibrium

Markers far apart or on different chromosomes will generally be independently distributed in a homogeneous population, a situation known as *linkage equilibrium*. Associations at neighboring loci, known as *linkage disequilibrium* (LD) (Ott, 1999; Weiss, 1993) can arise when new mutations appear or populations with differing allele frequencies mix. Such associations among alleles will gradually decay over generations at a rate that depends upon the physical distance between loci and underlies the most widely used approach to mapping disease genes: association studies. LD can result in associations between marker alleles and alleles of a causal variant, such that certain marker alleles will be present more often in affected individuals than in a random sample of individuals from the population. Thus, one can estimate the association between specific alleles and disease status; when an association is observed, the locus may be causal, or more likely be in LD with a putative disease locus. Another important mechanism for the appearance of LD at unlinked loci, even on different chromosomes, is *population stratification*, which can lead to spurious associations (discussed further below).

Investigating the Potential Genetic Basis of Disease

Familial Aggregation and Recurrence Risks

Often the first step to investigating a possible genetic etiology of a disease is establishing whether it "runs in the family" (*familial aggregation*). Clustering within families suggests genetic and/or shared environmental risk factors for a disease. A simple evaluation of aggregation may entail a standard case-control comparison of family history of a disease. Based on such a design, it can be determined whether cases are more likely than controls to have one or more close relatives affected. There are many ways to define family history and more sophisticated tests to assess familial aggregation. Some permit calculation of the recurrence risk of disease, the ratio of the risk for an individual with an affected relative to the baseline risk of disease in the general population. Recurrence risks are most commonly calculated for siblings, with larger values supporting a genetic basis for disease. However, familial aggregation and increased recurrence risks do not themselves indicate a genetic etiology, as they could be due in part to sharing of environmental factors for disease.

Twin Studies and Heritability

The study of twins also permits evaluation of genetic involvement in disease. If a disease occurs more often among both identical (*monozygotic*, MZ) twins than both fraternal (*dizygotic*,

DZ) twins or siblings, then genetics likely play a role (assuming environmental factors are shared equally by MZ and DZ pairs). One can calculate *narrow sense heritability*, which is the proportion of phenotypic variance attributable to additive genetic effects (excluding dominance, epistasis, and gene-environment interactions). A classic example is a study of 4,840 male twin pairs tracked by the Swedish Twin Registry and Swedish Cancer Registry (Gronberg, Damber, & Damber, 1994). The authors observed 458 occurrences of prostate cancer. Sixteen MZ but only six DZ twin pairs were concordant for prostate cancer, yielding proband concordance rates of 0.192 and 0.043 respectively, suggesting that genetics may play a substantial role in the development of the disease. In a similar study of Finnish twins, 12,941 same-sexed twin pairs were investigated for the incidence of primary cancers. Based on 1,613 malignant neoplasms, the authors found that the co-twins of MZ twins with cancer had a 50% greater risk of developing cancer themselves than dizygotic co-twins (Verkasalo, Kaprio, Koskenvuo, & Pukkala, 1999). Nevertheless, a subsequent, much larger study of 44,768 of Swedish, Danish, and Finnish twins (Lichtenstein et al., 2000; Mucci, Hjelmborg, Harris, & et al., 2016) concluded that environmental factors may have a larger effect on most cancer sites than germline genetics.

In addition to estimating heritability, twin studies can be used in other ways. Comparison of twins reared apart vs. twins reared together allows one to estimate both genetic and shared environmental effects, thereby overcoming one of the most important challenges to the classical twin design (Heston, 1966). MZ twin pairs discordant for disease can also be viewed as genetically-matched pairs for study of environmental risk factors, and also provides a different perspective on risk factors by asking each about both their own and their cotwin's histories (Hamilton & Mack, 2000). For disease-discordant twins, one might compare the case-spouse to the cotwin-spouse. Shared environmental factors can be compared between disease-concordant and discordant twin pairs, treating the pair as the unit of observation: factors that are more common among concordant pairs would be interpreted as risk factors, while differences in relative risks between MZ and DZ pairs would be evidence of an interaction effect. Twin-family (Laitinen, Rasanen, Kaprio, Koskenvuo, & Laitinen, 1998) and cotwin-control (Duffy, Mitchell, & Martin, 1998) designs allow one to study even more complex hypotheses.

Segregation Analysis

Segregation analysis also does not require DNA from study subjects as it is aimed at determining the mode of inheritance for a disease. It seeks to establish the potential genetic basis of disease, whether a single or multiple loci are involved in disease, and whether they act in a dominant, recessive, or co-dominant fashion. Segregation analysis also estimates allele frequencies and penetrances by fitting parametric models to familial phenotype data using maximum likelihood.

To establish a genetic etiology with segregation analysis, one tests the various Mendelian inheritance models against a more general alternative that might mimic environmental transmission, such as the *ousiotype* model in which risk also clusters into three categories like genotypes, but their probabilities are independent across generations (Hopper, 1989). A challenge to segregation analysis of disease traits (and particularly rare diseases) is that families are not generally randomly sampled from the population. More often, they are ascertained through one or more affected *probands*. One must account for this ascertainment through a likelihood calculation that conditions on family selection. This requires that families be ascertained according to a well-defined statistical sampling scheme, which may not be possible when families are derived from genetic counseling clinics, for example.

In addition to testing the major gene inheritance models, segregation analysis generally

tests for *polygenic* models. Here, the disease is assumed to be caused by many genes, each of which has small effects, so one can assume that that their aggregate effect is normally distributed in the population. Then each individual's polygene has an expectation midway between that of the two parents and a variance half that of the population variance. To compare major gene and polygenic models, both are nested within the general mixed model that includes both.

Linkage Analysis

Given evidence of a genetic basis of a disease, the location of the relevant genes may be sought using linkage analysis. This approach searches for genetic markers that are transmitted through families in a manner closely paralleling the transmission of the disease. The approach requires DNA from affected pairs of relatives or extended pedigrees, and relies on the phenomenon of chromosomal recombination.

Parametric linkage models require specification of a genetic inheritance model (allele frequencies and penetrances), often obtained from a prior segregation analysis. These models use likelihoods that involve two loci, the marker locus and an unobserved disease locus, separated by a recombination fraction θ that will be estimated. To avoid the difficulties of ascertainment correction (since heavily loaded pedigrees with many diseased individuals are generally used), a conditional likelihood is formed for the probability of the marker data given the phenotype data; conditioning on the phenotypes renders the method of ascertainment irrelevant. The results from such models are given in terms of the *lod score*, which provides evidence in support of or against linkage. In the latter case, the region around the marker locus might be excluded from further consideration as containing a causal variant.

Nonparametric or model-free linkage analysis uses affected pairs of relatives, generally sibling pairs. It does not require specification of a genetic inheritance model or extensive likelihood calculations. Instead, one searches for markers that are shared by affected pairs IBD more frequently than expected by chance. Sibling pairs, as noted above, are expected to share 50% of their alleles IBD, so a greater percentage would be evidence of linkage. One can apply a test of the 50% value and estimate the actual sharing using simple methods for proportions (Ch. xx), but the approach does not provide an estimate of the recombination fraction.

The heyday of linkage analysis was the last two decades of the 20th century, when dense panels of hundreds of highly polymorphic multi-allelic markers became available that made scans across the genome possible. Once a region was identified as possibly containing a disease-causing genetic variant, additional flanking markers would be tested and *multipoint linkage analysis* methods (Lander & Green, 1987) would be used to further localize the variant. Nevertheless, the resolution of linkage analysis for most realistic sample sizes was only about 1% recombination or 1 centiMorgan ($\theta = 0.01$), corresponding to about one million base pairs. This is a large region for trying to identify a potential causal genetic variant.

Association Studies

Association studies can be more powerful than linkage analyses and can delineate more manageable regions potentially harboring disease-causing variants. As a result, association studies have displaced linkage as the primary approach in genetic epidemiology. The aim of association studies is to provide evidence for association between markers and disease. These studies can be undertaken among unrelated individuals or within families.

Unrelated Individuals

Most commonly, association studies use cohort or case-control designs to compare marker allele or genotype frequencies between a group of unrelated affected individuals and a group of unrelated unaffected individuals. Association studies can also evaluate continuous traits, studying entire groups of subjects or selecting subjects at the extremes of the trait distribution to increase the statistical information (power and precision) (Huang & Lin, 2007). Whatever the phenotype, study subjects should be representative of their respective source populations to allow for generalization of results and to help address issues ot potential population stratification bias (Wacholder, McLaughlin, Silverman, & Mandel, 1992). In a case-control study, this implies that controls should be individuals who, if diseased, would have been cases (see Chapter 8). Controls are also commonly selected to match cases with respect to ethnicity, age, sex, and other potential confounders. Nevertheless, many association studies have been successful without overly rigorous control selection. In fact, due to the high cost of subject recruitment and genotyping, there is a growing movement toward using genotype information among controls recruited into previous studies or large publicly-available population surveys like the UK Biobank (Luca et al., 2008; S. G. Thompson & Willeit, 2015). The potential bias arising from using 'convenience' controls is tempered by the low potential for important confounding or selection effects—since external factors seldom affect the genotype "assigned" at conception. There is also low potential for measurement error in SNP genotyping, lack of recall bias when studying inherited variants, large sample sizes, and potential for rigorous replication of findings.

In addition, using such controls can result in population stratification bias—confounding of associations between particular genetic characteristics and the outcome by the rest of the genome –and overdispersion of test statistics, due to case-control differences in genetic ancestry (D. C. Thomas & Witte, 2002). This can be addressed analytically using a large panel of genetic markers to characterize and adjust for genetic ancestry (Devlin & Roeder, 1999; A.L. Price et al., 2006; J. K. Pritchard & Rosenberg, 1999; Zhu, Li, Cooper, & Elston, 2008).

Family-based Association Studies

Family-based association study designs have the appeal of providing protection from confounding by population stratification because all the subjects within a family are likely to have similar ancestry. There are two basic modes of this design, one comparing affected and unaffected relatives as matched case-control sets, and one leveraging transmission of alleles from parents to affected offspring. The former mode is straightforward when using full siblings as controls, as matched sets will have the same parents and hence identical ancestries. However, if the marker under analysis is not the causal genetic variant, then the results must be interpreted as referring to linkage and association. Furthermore, if multiple cases or multiple controls from the same sibship are included, then the usual conditional likelihood for matched sets is no longer valid because transmissions to each subject are not independent conditional on their allele sharing. In this situation, a robust variance estimator must be used (Siegmund & Gauderman, 2001). If more distant relatives are included, then their ancestries may not be identical, so protection from confounding is incomplete.

The second family-based association study design compares the alleles transmitted from parents to affected offspring with those not transmitted, which can be done using methods for matched-pair analysis (Ch. xx), such as a form of McNemar's test of alleles known as the *transmission-disequilibrium test* (Spielman, McGinnis, & Ewens, 1993). In this design, only the case and parents are genotyped and the disease status of the parents is not used. Equivalently, the design can be thought of as a 1:3 matched case-control comparison of the case to the "pseudosibs" carrying the other possible genotypes the case could have inherited.

Studying families can be statistically less efficient than studying unrelated subjects (Witte, Gauderman, & Thomas, 1999). This is because the genetic similarity among family members can result in fewer discordant genotypes sets for analyzing main effects than in studies of unrelated individuals. On the other hand, for analyzing gene-environment interactions, power or precision may be enhanced, essentially because the most informative comparisons are genotype-concordant exposure-discordant pairs, which tend to be enriched despite their potential "overmatching" on exposure precisely because of their greater concordance in genotype. Moreover, selecting unrelated cases with a positive family history of disease can increase the information in association studies (Antoniou & Easton, 2003).

Analyzing Associations

Whatever the design, the association between genetic variants and disease is most commonly evaluated using a trend test across the number of variant alleles. This allelic trend test is relatively robust to model misspecification (e.g., if the true mode of inheritance is recessive or dominant rather than additive). Investigators generally use a regression model for the variant-disease association that adjusts for confounders. An unbiased, systematic association has at least two interpretations (Lander & Schork, 1994). First, the disease-associated variant might be the susceptibility allele itself. Second, the associated allele might be in LD with the susceptibility allele at the disease locus. In the first case, the marker and disease locus are one in the same, so LD is complete and the association is expected to occur in all populations harboring the allele. In the second case, the marker and disease locus are very close to each other, but differences in LD may lead to different associations across populations. For this reason, association studies can narrow the region potentially containing a disease-causing marker. Multiethnic studies provide a valuable approach to distinguishing these possible explanations by exploiting differences in LD structure across populations.

Genome-wide Association Studies

Most early association studies among unrelated individuals focused on specific *candidate genes* with suspected functionality on a pathway to the phenotype. In the late 1990s, however, emerging microarray chip technologies began to enable the interrogation of all the common variants across the genome for association with disease or quantitative traits. Investigators quickly realized that clusters of variants on the same chromosome (haplotypes) could effectively tag many other variants, so that it was really only necessary to genotype a modest number of variants at any locus to capture most of the common variation. The advent of the genomics era in genetic epidemiology can be dated to a seminal paper in *Science* by Risch and Merikangas (1996), in which they predicted it would soon become possible to search for disease genes in an agnostic fashion by testing hundreds of thousands of such 'tag SNPs' for association with disease.

Two major advances soon made this vision a reality. Further improvements in genotyping technologies brought down the cost of assaying SNPs genome-wide to a few thousand dollars per sample (a trend mirroring Moore's Law in computing that by 2018 had lowered the cost for millions of variants to less than one hundred dollars). Then, with the creation of the International Haplotype Mapping Project (HapMap) (Gibbs et al., 2003), investigators began systematically cataloging haplotype tagging SNPs across the entire genome in multiple populations (Daly, Rioux, Schaffner, Hudson, & Lander, 2001; Frazer et al., 2007; Gabriel et al., 2002; HapMap, 2003, 2005). Less than a decade after Risch and Merikangas' paper, the first genome-wide association study (GWAS) was published (Klein et al., 2005) along with two confirmatory studies (Edwards et al., 2005; Haines et al., 2005) reporting an association

of age-related macular degeneration with the complement factor H gene *CFH*. This association helped localize a gene within a previously known linkage region, demonstrating the power of GWAS.

In the early days of GWAS, the high cost of genotyping motivated the use of multi-stage designs (Satagopan, Verbel, Venkatraman, Offit, & Begg, 2002). In the first stage, part of the sample was genotyped using a GWAS array. In the second stage, the remaining samples were genotyped for the most strongly associated markers using a less expensive custom panel. Information from the two stages was then combined in a final analysis. With declining genotyping costs for GWAS compared with custom panels and the recognition that genome-wide data are useful for many other purposes, such as testing for gene-environment and gene-gene interactions, most studies conducted at present aim to obtain GWAS data on an entire sample (D. C. Thomas et al., 2009).

The number of GWAS undertaken has expanded rapidly since 2005. The association results from these studies with $p < 1x10^{-5}$ have been catalogued by the National Human Genome Research Institute and European Molecular Biology Institute (MacArthur et al., 2017). As of November 2017, the searchable database (http://www.ebi.ac.uk/gwas/home) contained 53,020 unique SNP-trait associations from 3,197 publications. The vast majority are weak associations, with relative risks between 1.1 and 1.3.

Very large sample sizes are required to study associations with ever-smaller effect sizes, which has motivated the creation of large consortia, often comprising tens of thousands of cases and controls. Although sharing of raw data at the individual level across a consortium, each with their own consent and data sharing rules, can be problematic, a joint analysis can often be conducted by meta-analysis of summary statistics, at least for genetic main effects.

To date, for most complex diseases, the proportion of the estimated heritability explained by SNPs detected in GWAS is somewhat limited, as is the area under the curve (AUC) of receiver operating characteristic (ROC) curves for these SNPs. For example, even for height — a highly heritable trait that is easily and accurately measured — an early meta-analysis of almost 200,000 individuals found 180 associated loci, but together they accounted for only 10% of the phenotypic variation (Lango Allen et al., 2010). The authors estimated that as many as 600 loci with similar effect sizes might be detectable with sample sizes of 500,000 subjects, but the explained variation would still be only 16%. Chatterjee and Park (2011) described the AUCs for available genetic loci of complex diseases, with and without including family history and non-genetic risk factors, and for most cancers they were under 0.65 (with 0.5 being the expected value for a risk index that is no better than chance). Inclusion of SNPs predicted to be discovered in the future only increased the estimates by a modest degree. Only for Crohn's disease and type I diabetes were AUCs greater than 0.75. That said, there is a growing appreciation that combining GWAS risk variants into a polygenic risk score might offer some "translational" value for predicting risk to guide screening and prevention strategies or for stratification in therapeutic studies and treatment planning (F. Dudbridge, Pashayan, & Yang, 2017).

There has been considerable discussion about the "hidden" or "missing" heritability that remains to be explained (Eichler et al., 2010). With regard to what is "hidden", a much larger proportion of heritability can be explained when looking at all variants genotyped or tagged by GWAS arrays (e.g., up to 80% for height) (Lango Allen et al., 2010; Yang et al., 2010). For example, Purcell et al. (2009) showed that risk indices comprising increasing numbers of the "top" SNPs (up to as much as half the genome) produced better and better prediction of the risk of schizophrenia and bipolar disorder (but not for heart disease, rheumatoid arthritis, hypertension, Crohn's disease, or diabetes). "Missing" heritability might be explained with larger studies focused on rare variants with larger effect sizes that are not covered by current

GWAS panels, or considering gene-gene or gene-environment interactions. Whole genome sequencing may provide some of these clues (discussed further below).

Imputation of Genetic Markers

Most associations discovered via GWAS are unlikely to be directly causal, because only a limited fraction of the genetic variation in the genome is directly measured. Marker associations are instead expected to reflect indirect relationships with nearby causal variants in LD. With the availability of extensive reference panels of variation from projects such as the Haplotype Reference Consortium (Iglesias et al., 2017), it has become possible to infer genotypes for most variants in the genome with >1% minor allele frequency (MAF) by using imputation techniques (Y. Li, Willer, Ding, Scheet, & Abecasis, 2010; Marchini & Howie, 2010). Although some programs provide the most likely genotype at each untyped locus, the use of best guess genotypes in association analysis fails to account for the uncertainty, thereby leading to biased tests. A more appropriate procedure is therefore to use the estimated genotype probabilities or (under an additive model) the expected "geneotype dosage" as a continuous variable in regression models (Hu & Lin, 2010). Prior to imputation, quality control of genotyping data is essential for valid results (Pluzhnikov et al., 2010). Standard practice involves eliminating samples with undue numbers of loci that fail certain quality control filters and eliminating markers with low genotype call rates and departures from Hardy-Weinberg equilibrium.

Population Stratification

In the analysis of association study data, one must account for potential confounding due to population stratification. If variant allele frequencies and disease risks vary across ancestral populations, genetic associations at a particular locus can be confounded by the genetic associations in the rest of the genome. Such confounding can lead to bias in statistical results, even within apparently homogeneous populations. Various methods have been developed to account for such population stratification using the entire distribution of associations. The most widely used method today is to adjust for principal components of genetic ancestry (A. L. Price et al., 2006). Alternatives are to adjust for estimates of the proportion of the genome inherited from different ancestries (J K Pritchard, Stephens, Rosenberg, & Donnelly, 2000), or to use mixed models that combine features of both approaches and can be applied to family-based or case-control studies (Zhu et al., 2008). A typical diagnostic used to assess potential population stratification is to plot the observed cumulative distribution of *p*-values against its expected uniform distribution (a Q-Q plot) and verify that its median is close to ½ ($\lambda = 1$). When using ancestry estimation, it may be important to adjust not just for global ancestry, but also for the local ancestry in each region of the genome, particularly in admixed populations like Hispanics or African Americans (Zhang & Stram, 2014).

Statistical Testing for GWAS Studies

To obtain valid tests in GWAS analyses, one must address the issue of multiple comparisons arising from evaluating millions of associations. Pe'er et al. (2008) have estimated that for modern high-density genotyping chips with imputation to reference panels or next generation sequencing, which may allow one to assess tens of millions of genetic variants, the linkage disequilibrium structure translates to approximately a million effectively independent comparisons in Europeans or about two million in Africans. The simplest approach is then to use a Bonferroni correction, in which α is divided by the number of effectively independent tests performed, e.g., with $\alpha = .05$ and a million independent associations, the new cutpoint is

$0.05/1{,}000{,}000 = 5\text{x}10^{-8}$. More refined methods have better performance and more scientific relevance, however (e.g., (Efron & Hastie, 2016), Ch. 15). Some GWAS also calculate the false discovery rate (FDR) or use permutation testing to assess the strength of associations (Storey & Tibshirani, 2003; D. C. Thomas & Clayton, 2004; Wacholder, Chanock, Garcia-Closas, El Ghormli, & Rothman, 2004b). Note that while adhering to strict "significance" cut-points may help address issues of multiple comparisons, the cut-points are somewhat arbitrary and do not reflect the potential clinical or biological importance of associations (Witte, Elston, & Schork, 1996), nor do they account for the costs of false negatives. A case in point is the increase in heritability explained when considering all variants measured by a GWAS array, and not just those reaching genomewide significance. Furthermore, an important aspect of GWAS findings is not simply "statistical significance" but also their validity, as examined by replication studies on independent samples (Ioannidis et al., 2008; Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis, 2001; Ioannidis, Thomas, & Daly, 2009; Kraft, Zeggini, & Ioannidis, 2009).

Recent Advances in Association Methods

Next Generation Sequencing

GWAS usually assay known genetic variants with MAF greater than 5%. Other common SNPs are well tagged by SNPs on GWAS arrays, as are "uncommon" variants with MAF from 1–5%. To measure rare variants (<1% frequency), one can attempt to impute them, add custom genotyping content to a GWAS array, or undertake next generation sequencing (NGS). NGS also allows for the measurement of genetic variants that are not known or not amenable to measurement or imputation by genotyping, such as *copy number variants* (CNVs). Thus, the appeal of NGS is the ability to discover *all* the variants in a gene, in the entire exome, or even the entire genome, and not just the common variants. A popular hypothesis is that these rare variants may have larger effect sizes than common variants (Bodmer & Bonilla, 2008), but note that some of the purported evidence for this hypothesis may be due simply to the lower power for rare variants, so that only those with large effect sizes are detectable!

To perform NGS, one randomly breaks samples of DNA from each chromosome into many small overlapping fragments, such that any given locus is spanned by a random number of fragments. These fragments are then sequenced and aligned against a reference sequence to report the number of copies of each nucleotide base at each location. A sample that is heterozygous at a given location may or may not be correctly called as such; the call depends upon whether both alleles are covered by one or more fragments, which in turn depends upon the *depth of sequencing* (the average number of copies across the genome). Because sequencing is less than perfect, one might want to dismiss an observation of only one or two copies of the variant allele as a false positive, so that greater depth of sequencing would be needed.

Due to the expense of next generation sequencing, there is a trade-off between average depth of sequencing and the number of samples that can be sequenced. If the goal is to discover rare alleles (which would occur predominately in the heterozygous state), then the number of samples tested is far more important than depth, suggesting one might prefer large-scale low-coverage studies (Ionita-Laza & Laird, 2010; Y. Li, Sidore, Kang, Boehnke, & Abecasis, 2011). An alternative is to combine or 'pool' DNA samples. Here one could use either a two-stage design, initially using pooling to identify sets of samples containing variants of interest followed by individual sequencing to identify specific samples containing those variants (Xu, Wu, Zhang, Shen, & Deng, 2017), or directly use a test of association comparing estimated allele frequencies in case and control pools (Liang, Thomas, & Conti, 2012).

Rare Variant Analyses

Conventional approaches to assessing association are inefficient when studying rare variants. This has given rise to numerous novel statistical tests for rare variant association, generally falling into two main categories: burden tests and variance component kernel machine tests. Burden tests aim not at testing association with individual variants but rather with the number of variants or some weighting of their aggregate effect in a gene region (Bansal, Libiger, Torkamani, & Schork, 2010; Basu & Pan, 2011). Since their introduction in 2011, a more widely used approach has been the variance component sequence kernel association test (SKAT), which tests for some measure of genetic similarity in a region between pairs of subjects with similar phenotypes (e.g., greater genetic similarity between case-case than case-control pairs) (Lee, Wu, & Lin, 2012). Burden approaches are more powerful if all of the rare variants under study have association effects in the same direction. Conversely, variance component approaches are more powerful when some rare variants are positively associated and others are negatively associated. One can also use an approach that compromises between the burden and variance component approaches (SKAT-O) (Seunggeun Lee et al., 2012).

Another important development in the study of rare variants has been a resurgence of interest in family-based designs. These have three advantages: first, by focusing on families with multiple cases, the frequency of causal variants is likely to be enhanced; second, they allow for more powerful tests of causality by identifying variants that co-segregate within families along with the inheritance of the disease; third, they provide a way to get lots of "free" genotyping by imputing the genotypes of untested family members based on those of their close relatives. Variations on SKAT and other rare variant tests have been developed for family data (Schifano et al., 2012). A particularly appealing design is a two-stage family-based design in which a subset of the most informative family members (based on their disease status and prior linkage or GWAS markers) is sequenced and used to prioritize a subset of variants most likely to be causal for testing in an independent sample (Duncan C Thomas, Yang, & Yang, 2013). This contrasts with a "two phase" design for case-control samples of unrelated individuals (Breslow & Chatterjee, 1999), in which a subsample of individuals is chosen for sequencing based on their disease status and associated markers, and then used to impute genotypes for the remaining sample in a joint analysis of the main study and subsample.

Evaluating Multiple Risk Factors

Interactions

Following an initial scan for main effects of SNPs from GWAS or rare variants from NGS, much more remains to be explored. One possibility is that there could be gene-environment or gene-gene interaction effects that do not involve markers producing important marginal effects. (We caution that we are here following the traditional genetic use of "interaction" to mean modification of a multiplicative measure, rather than some biological interaction; see Chapter 23). A challenge to evaluating interaction is that the number of possible interactions is much larger than the number of main effects. For example, a GWAS of 1 million SNPs will have a half a trillion possible pairwise interactions. A simple Bonferroni correction for multiple comparisons would thus require an alpha level of 1×10^{-13}, and interaction tests require much larger sample sizes than main effects even at the same alpha level (Marchini, Donnelly, & Cardon, 2005).

The study of interactions can be enhanced by "case-only" analyses (see Chapter 8) that assume independence of the interacting factors in the source population. For example, a re-analysis of cleft palate data obtained narrower confidence limits for the interaction between

smoking and the *TGFα* gene (an odds ratio (OR) of 5.14 with 95% confidence limits (CL) of 1.68 and 15.7 for the case-only analysis compared with an OR of 6.57 with CL of 1.72 and 25.0 for the case-control analysis), equivalent to a 30% reduction in sample size required for the same precision (Umbach & Weinberg, 1997). Case-only analyses are, however, biased if the independence assumption is violated, as might arise due to LD among nearby pairs of variants, population stratification, or behavioral factors that induce an association between genes and environmental factors. Unlike the case-control design, the case-only design can only measure interaction on a multiplicative scale (see below).

To overcome these concerns, various staged and hybrid approaches have been introduced (Evans, Marchini, Morris, & Cardon, 2006; Kooperberg & Leblanc, 2008; D. Li & Conti, 2009; Mukherjee & Chatterjee, 2008; Murcray, Lewinger, Conti, Thomas, & Gauderman, 2011). Although power will still be much lower for interactions than for main effects, these methods generally yield much better power than a simple exhaustive search (Cornelis et al., 2012; Mukherjee, Ahn, Gruber, & Chatterjee, 2012). Moreover, for future studies of gene-environment interactions, it is essential that investigators planning new GWAS studies design them to have appropriate environmental measurements and population-based sampling schemes. For example, the NIH post-GWAS "GAME-ON" initiative aimed at synthesizing all available data on five cancer sites, replicating findings, and characterizing genetic risks and their modification by environmental exposures (Fehringer et al., 2016) has been limited by many of the available studies not having collected any environmental exposure data. Even those studies that did collect such data did so in a highly variable manner; the data range from very crude to very detailed, such that they often require considerable efforts at harmonization across studies (Bookman et al., 2011).

It has long been recognized that any statement about interaction is necessarily scale dependent. The dominance of the logistic model and odds ratios for estimating associations has led to a focus of interaction measuring departures from the multiplicative model, but it has also been recognized that the public health concept of "synergy" is often better captured by departure from the additive model and that the simplest biological models of interaction produce departures from additivity, not multiplicativity (Rothman, Greenland, & Walker, 1980). For example, while radiation exposure has been shown to have a greater-than-multiplicative effect with the *ATM* gene on the risk of second breast cancers following radiotherapy (J. L. Bernstein et al., 2010), its interaction with *BRCA1* and *BRCA2* was not shown to be different from multiplicative although it may be somewhat greater than additive (Jonine L. Bernstein et al., 2013). Thus, while carriers of these mutations are at considerably increased baseline risk, radiation exposure could still further enhance their risk.

Pathways

Another way to consider multiple risk factors is to recognize that SNPs that individually fail to reach genome-wide significance may jointly contribute to a common pathway. A variety of methods have been developed for identifying subsets of genes in known pathways that are collectively over-represented among the top GWAS associations (D. C. Thomas, 2012; Wang, Li, & Hakonarson, 2010). The first and perhaps best known method is Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). GSEA and other approaches entail some way of combining the study data, say from a GWAS or eQTL study, with external information. The external information is typically highly structured using expert systems approaches to organizing information in a computable format, known as an "ontology." The Gene Ontology (Gene Ontology et al., 2013) is an example of such a database aimed at characterizing gene products in terms of cellular component, biological process, and molecular function. Entries in the Gene

Ontology are manually curated by a small army of biological experts using phylogenetic relationships amongst members of a gene family across species. Other ontologies include the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2008) and Ingenuity Pathway Analysis (Kramer, Green, Pollard, & Tugendreich, 2014). Our understanding of pathways is continually curated into various databases in particular biological domains (e.g., genomic, pathway, or functional). Other systems, like the GRAIL program (Raychaudhuri et al., 2009), use statistical techniques based on expert systems to mine the database of PubMed abstracts to identify functional relationships among a set of genes.

Hierarchical Modeling

Yet another approach to incorporating additional information in genetic epidemiology is Bayesian hierarchical modeling (Capanu et al., 2011; Capanu et al., 2008; D V Conti et al., 2009; D. V. Conti & Witte, 2003; Hung et al., 2007; Lewinger, Conti, Baurley, Triche, & Thomas, 2007; Wilson, Baurley, Thomas, & Conti, 2010). Unlike other methods that require prior specification of the importance of external information, hierarchical modeling uses a higher-level regression model to infer from the study data the utility of such information (e.g., annotations) across the entire set of associations under consideration. The inclusion of relatively uninformative annotations in the second-level model produces very little loss of power because such variables received little or no weight, whereas including truly informative annotations greatly improves power (Quintana & Conti, 2013; Quintana et al., 2012). In contrast, standard methods with pre-specified weights lead to substantial loss of power when the weights are incorrectly specified. For example, Minelli et al. (2013) compared the weights for 15 types of knowledge about SNPs elicited from 10 experts using a Delphi survey versus empirical estimates of their predictive values for 7 diseases; they found some agreement but also some disagreement (e.g., the empirical evidence did not support the importance of a gene encoding a protein in a pathway or interactions relevant to a trait that were ascribed by experts). A companion paper (J. R. Thompson et al., 2013), showed how both kinds of external information could be used to estimate the probability of association in a hierarchical modeling framework. For example, data from an epidemiologic study of gene-environment interactions in asthma were combined with an experimental challenge study using model air pollutants in a panel of allergic subjects in a multivariate hierarchical model, along with external pathway information from Ingenuity Pathway Analysis (R. Li, Conti, Diaz-Sanchez, Gilliland, & Thomas, 2013).

Mendelian Randomization

In 2004, George Davey Smith and Shan Ebrahim (2004) reprinted a 1986 letter to the editor of *The Lancet* by Martijn Katan (1986) in the *International Journal of Epidemiology.* It proposed a way to evaluate the causality of the observed association between serum cholesterol and cancer by relating both variables to the *ApoE* genotype. Although the approach was never performed, it appears to be an early reference in the medical literature to a technique long used in econometrics (S. Wright, 1928) known as "instrumental variables." In the context of genetic epidemiology, the basic idea is to indirectly analyze an association between some exposure 'biomarker' (e.g., serum cholesterol) and disease by examining the extent to which a genotype that is strongly related to the biomarker is also related to disease risk solely through that intermediate variable. Here, genotype functions as the instrumental variable (Greenland, 2000).

This approach, now known as Mendelian randomization (MR), attempts to address two well-known problems in directly testing associations of a biomarker with a disease or trait in case-control or cohort studies. The first is reverse causation, whereby the disease process or its treatment alters the biomarker, rather than the biomarker affecting the risk of disease. This

problem is more severe in retrospective case-control studies, but could affect cohort studies using stored biospecimens if the disease has a long pre-symptomatic period. The second problem is uncontrolled confounding, a near universal problem in observational epidemiology. Because constitutional genotypes are immutable, there is no possibility for reverse causation, and because genes are transmitted randomly from parents to offspring, confounding is avoided, at least within families, although spurious gene-biomarker or gene-disease associations can arise in comparisons across populations. MR does require certain key assumptions (Didelez & Sheehan, 2007), the most important being that there is no pathway from the genetic marker to disease other than through the biomarker (no "pleiotropic" effects). For any given gene, this can be essentially impossible to know with certainty.

There has been growing enthusiasm for applying MR in the context of GWAS, scanning first for a subset of genes that are predictive of the biomarker and then analyzing their association with disease. This can be done using separate samples for testing the two associations ("two-sample MR") and in the context of large-scale consortia for greater power (Burgess, Butterworth, & Thompson, 2013). However, the no-pleiotropy assumption becomes less convincing as many variants are included for which little is known about function. As a result, methods such as weighted median regression (Bowden, Davey Smith, Haycock, & Burgess, 2016) have been developed that only require half of the variants to be valid instrumental variables. In GWASs, the expectation is that few of the variants would actually be causally related to a phenotype, so even if some might have pleiotropic effects through mechanisms unrelated to the biomarker under study and hence would not be valid instrumental variables, it is unlikely that the majority of them would be invalid. Thus, the median of the ratios of the regressions of individual SNPs associations on the biomarker and the phenotype should be a valid estimator of the effect of the biomarker on the phenotype. Greater precision can be obtained by using a weighted estimated of the median, giving more weigh to the more precise estimates of the ratios.

Ethical, Legal, and Social Issues

The field of genetic epidemiology raises a number of important ethical, legal, and social issues (ELSI). These include the potential value of genetic testing, the possibility of genetic discrimination, reporting research findings to study subjects and their family members, and incorporating genetics into treatment and prevention. Results of genetic epidemiology research can thus impact individuals who carry markers for disease, their families, and society as a whole.

The growing number of discoveries from genetic epidemiologic research and decreases in genotyping and sequencing costs have resulted in increasing availability of genetic testing. Such tests range from those ordered by a physician for high risk disease genes to those marketed direct-to-consumer for modest or low risk variants that span the entire genome (Kaye, 2008). Clinical genetic tests are generally coupled with genetic counseling to help individuals understand the risk of disease associated with carrying disease markers. Such tests can, however, also detect incidental findings such as genetic mutations from sequencing that are not directly related to the disease originally under study. Determining how to report such results to individuals has important ELSI implications.

In addition to informing treatment decisions, genetic testing has potential implications for both primary and secondary prevention (Rebbeck, 2014; D. C. Thomas, 2017). For example, should carriers of a mutation that increases susceptibility to a particular chemical be barred from occupational exposure or should they have a right to choose once properly informed? Should disease screening programs be designed to target individuals with high genetic risk scores?

Some tests include information about the genetic basis of drug response, which is now

being studied using GWAS (Rieder, Livingston, Stanaway, & Nickerson, 2008). Pharmacogenomics has important near-term implications for what drug and dose an individual should receive. For example, GWAS confirmed that variants in the gene *CYP2C9* impact metabolism of the anticoagulant drug warfarin (Takeuchi et al., 2009); such findings suggest that individuals who carry the *CYP2C9* variant could be prescribed lower doses of warfarin, reducing potential side effects of severe bleeding and unnecessary health-care costs. In light of such results, precision medicine is commonly touted as a major potential benefit of pharmacogenomics and GWAS—albeit with some reservations (Nebert, Zhang, & Vesell, 2008; Omenn, 2009).

Direct-to-consumer testing generally offers individuals the opportunity to have variation in their genomes assayed with the same genotyping arrays used in GWAS. Results from these assays are returned to the consumer along with additional information such as their genetic ancestry and estimated health risks. Understanding the implications of risk is challenging when the effect estimates are modest. This raises the question of how to regulate recommendations about behavior modification from direct-to-consumer testing services to ensure they are based on sound science. Concerns here have led the Food and Drug Administration to restrict what information can be provided by direct-to-consumer genetic testing companies to their customers. However, the Food and Drug Administration has approved direct-to-consumer testing of high risk *BRCA* mutations, which raises a number of concerns about overtesting and false assurances from negative results in low risk populaitons (Gill, Obley, & Prasad, 2018).

There are also a number of ELSI issues surrounding altering disease genes to treat or prevent disease. Gene therapy—whereby genes are therapeutically inserted to an individual's cells—has had many failures until very recently. The gene editing technology, CRISPR-Cas9, shows promise for altering disease genes, including the ability to not only edit genes but also to impact expression without altering the structure of an existing gene (Cong et al., 2013; Mali et al., 2013).

New Directions in Genetic Epidemiology

Advances in genetic, genomic, and other omic technologies over the past decade have yielded stunning progress in our understanding of the biological basis of disease. An obvious manifestation is the success of GWAS at discovering thousands of unforeseen loci related to hundreds of diseases (Welter et al., 2014). While individually these risk loci typically have very small effect sizes, in aggregate they account for an increasing proportion of disease heritability (Eichler et al., 2010; Manolio et al., 2009; Zaitlen & Kraft, 2012). Moreover, we can now assess uncommon and rare variants that will help decipher an increasing proportion of the genetic basis of disease (Casey, Conti, Haile, & Duggan, 2013).

Gene Expression and GWAS

The study of disease etiology and prognosis benefits from the rich spectrum of potential biomarkers that are rapidly becoming available. These include gene expression, epigenetics, and proteomics. With regard to gene expression, arrays and RNA-Seq can be combined with GWAS data to undertake agnostic scans for expression quantitative trait loci (eQTL) and their role in disease (GTEx_Consortium, 2015; He et al., 2013; Mele et al., 2015). eQTLs are loci that influence expression in the same or nearby genes (*cis*) or anywhere in the genome (*trans*). A salient example is a result based on RNA samples from 175 individuals across 43 tissues as part of the Genotype-Tissue Expression (GTEx) project (GTEx_Consortium, 2015) highlighting tissue-specific and shared regulatory eQTLs and subsequently linking them to GWAS variants. One can impute expression levels from GTEx into genotyped GWAS data (Gamazon et al.,

2015), providing an important avenue for assessing the relation between gene expression and disease without having directly measured expression. Related methods are also now being applied to identify epigenetic effects of environmental exposures and their role in mediating gene expression and ultimately disease phenotypes (Cortessis et al., 2012). These methods include measuring high-density panels, making Epigenome-Wide Association Studies feasible (Rakyan, Down, Balding, & Beck, 2011). Proteomics techniques are just beginning to mature to the point of providing insight into the biological processes in disease etiology and ultimately sensitive early markers of disease or progression (Stunnenberg & Hubner, 2014).

Exposome and Microbiome

High throughput, highly sensitive, multiplex metabolomics methods such as mass spectrometry may provide advances for studying essentially all environmental exposures (the "Exposome") (Wild, 2012) and to facilitate agnostic Environment-Wide Association Studies (Patel, Bhattacharya, & Butte, 2010). There remain numerous methodological challenges before the EWAS concept can be considered a real companion to GWAS (Chadeau-Hyam et al., 2013). Unlike the genome, environments are dynamic and the appropriate critical periods of exposure for etiology may occur in the distant past, so current measurements taken may not reflect earlier exposures. To avoid the problem of reverse causation, a cohort or nested case-control design is needed. In GWAS, confounding by population substructure may be addressable by genomic control techniques, but there is no obvious parallel for controlling host and environmental confounders of exposures. The measurement of exposures is fraught with measurement error (e.g., temporal variability, instrument error, and identification of unknown chemicals). The advent of EWAS also opens the possibility of Gene-Environment-Wide Interaction Studies (GEWIS)−agnostic scanning of all possible gene-environment interactions (D. Thomas, 2010).

One key component of the exposome, disruptions of the incredible diversity of micro-organisms that inhabit the body, has been increasingly recognized as important in numerous chronic diseases (Cho & Blaser, 2012; Schwabe & Jobin, 2013). There is an emerging view that a host and microbiota form a 'super-organism' in a symbiotic relationship that plays a role in risk of disease by mediating immune response. Dramatic examples include improvement in symptoms of inflammatory bowel disease with a change in flora after antibiotic and probiotic use (Morgan et al., 2012), the use of fecal transplants from healthy subjects to cure *Clostridium difficile* infections (van Nood et al., 2013), and weight loss in obese subjects (Ridaura et al., 2013). Evidence is mounting to link tumor promotion in many cancer types to the effects of bacterial microbiota (Cho & Blaser, 2012; Gargano & Hughes, 2014; Schwabe & Jobin, 2013).

Although the gut microbiome is thought to be largely set by age three, exposures and interventions later in life, including diet, antibiotics, and other lifestyle characteristics may affect the structure of microbial communities (Arumugam et al., 2011; O'Sullivan et al., 2013; Pepper & Rosenfeld, 2012). Alterations of the microbiome in turn affect local and systemic host physiology and thus risk of disease and response to therapy. There is a dearth of sophisticated statistical methods for relating complex microbial community structure to the metabolome and disease (K. Li, Bihan, Yooseph, & Methe, 2012). Microbiome research raises many of the same methodological challenges as the exposome, and is further complicated by community-level effects like diversity and resilience and rarely occurring species. By jointly considering the microbiome and environmental exposures, investigators will have an opportunity to examine host-microbial (Kinross, Darzi, & Nicholson, 2011) and microbiome-metabolome (Ursell et al., 2014) interactions, topics that have been largely neglected, in part due to their staggering methodological complexity.

Integrative Genomics

All of the above omics measures should be analyzed in a comprehensive manner to develop a fuller understanding of the disease process, but the methodological challenges are numerous (Chadeau-Hyam et al., 2013). With the explosion of high-density omics platforms and the enthusiasm for agnostic association testing, genetic epidemiology has had to seriously confront the inference problems posed by high-dimensional data. Doing so is the ultimate goal of the emerging field of *integrative genomics* (Kristensen, Frigessi, & Borensen, 2014; Schadt et al., 2005).

A key challenge of high dimensional data is that the number of variables p is much larger than the number of observations n (the "$p > n$" problem). This issue arises in both the frequentist and Bayesian contexts and various approaches have been taken to address it. For frequentists, Bonferroni adjustment of acceptable Type 1 error rates and modifications thereof that allow for the correlation between tests (Conneely & Boehnke, 2007) have become conventional. In the gene expression field, the use of the FDR has become conventional, as the expectation is that there will be many noteworthy findings (Storey & Tibshirani, 2003). Bayesian versions of the FDR concept have also been proposed (Wacholder, Chanock, Garcia-Closas, El Ghormli, & Rothman, 2004a; Wakefield, 2007; Whittemore, 2007).

Beyond testing for associations one-at-a-time, the real challenge for integrative genomics is variable selection in model building. It has long been appreciated that the importance of variables selected by some stepwise or greedy algorithms (e.g., in a conventional regression model) cannot be interpreted at face value, nor can confidence intervals on the parameters of a "best" model. A variety of approaches have been taken to address this problem in a frequentist framework, such as splitting the analysis into a discovery and testing step, cross-validation, or penalized regression (e.g., LASSO (Tibshirani, 1996) or elastic net (Zou & Hastie, 2005)). By itself, however, LASSO does not provide a formal significance test, only a sparse selection of variables, although there have been important developments in frequentist inference (Lockhart, Taylor, Tibshirani, & Tibshirani, 2014; Meinshausen & Bühlmann, 2010).

Bayesian variable selection has also been a viable approach since the seminal work of George and colleagues on "stochastic search variable selection" (SSVS) using a "spike and slab" normal mixture model (George & Foster, 2000; George & McCulloch, 1993). Although not originally proposed for the $p>n$ problem, Bayesian methods of variable selection have recently been extended for that purpose (Bottolo et al., 2013; Bottolo & Richardson, 2010; Ročková & George, 2013). An example in the context of rare variant association modeling is the Bayesian Risk Index (Quintana, Berstein, Thomas, & Conti, 2011)—and extensions using Bayes model averaging in the Bayesian Model Uncertainty method (Quintana & Conti, 2013)—as well as incorporating external information as priors in the iBMU approach (Quintana et al., 2012). These methods allow valid inference on both the set of variables and the set of alternative models via Bayes factors.

Machine Learning

Another emerging theme in the systems biology and genomics literatures has been the development of machine learning approaches to "big data," e.g., as a nonparametric approach to discovering high-dimensional interactions (Greene, White, & Moore, 2008; McKinney, Reif, Ritchie, & Moore, 2006; Moore et al., 2006; Motsinger-Reif, Dudek, Hahn, & Ritchie, 2008; Romualdi et al., 2003). Such methods simply look for patterns in the data in an agnostic manner. While there have been some notable discoveries using such techniques, it remains somewhat unclear how useful they will be for building interpretable biologically-based models. Given the high-dimensionality of genome-scale problems, the use of parametric models, starting with

relatively simple linear main effects models and adding in two-way or higher-order interaction terms as needed, may be a more promising strategy for finding biologically interpretable and reproducible models. This strategy risks missing some complex interactions that could exist in biology (Moore, 2003); however, it is less clear whether such effects extensively carry over into epidemiologic data.

Conclusions

Genetic epidemiology has evolved at a rapidly increasing rate over the last two decades, mirroring and largely driven by technological advancements and "big science" collaborations. These trends are likely to continue, with further novel techniques developed to interrogate the genetic and environmental processes underlying disease. This expansion will undoubtedly improve our ability to relate different types of information in ever-larger sample sizes while exploiting the wealth of biological knowledge available in public databases. Hopefully, improving our understanding of the genetic and environmental causes of disease will result in personalized prevention and treatment strategies that improve the public's health.

REFERENCES

Ambrosone, C. B., Freudenheim, J. L., Graham, S., Marshall, J. R., Vena, J. E., Brasure, J. R., . . . Shields, P. G. (1996). Cigarette smoking, N-acetyltransferase 2 genetic polymorphisms, and breast cancer risk. *Jama, 276*(18), 1494-1501.

Antoniou, A. C., & Easton, D. F. (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol, 25*(3), 190-202.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., . . . Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature, 473*(7346), 174-180. doi:10.1038/nature09944

Bansal, V., Libiger, O., Torkamani, A., & Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet, 11*(11), 773-785. doi:nrg2867 [pii]

10.1038/nrg2867

Basu, S., & Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology, 35*(7), 606-619. doi:10.1002/gepi.20609

Bernstein, J. L., Haile, R. W., Stovall, M., Boice, J. D., Jr., Shore, R. E., Langholz, B., . . . Concannon, P. (2010). Radiation exposure, the ATM Gene, and contralateral breast cancer in the womens environmental cancer and radiation epidemiology study. *J Natl Cancer Inst, 102*(7), 475-483. doi:djq055 [pii]

10.1093/jnci/djq055

Bernstein, J. L., Thomas, D. C., Shore, R. E., Robson, M., Boice, J. D., Stovall, M., . . . Haile, R. W. (2013). Contralateral breast cancer after radiotherapy among BRCA1 and BRCA2 mutation carriers: A WECARE Study Report. *European journal of cancer (Oxford, England : 1990)*, Epub 4/18/13.

Bodmer, W., & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet, 40*(6), 695-701.

Bookman, E. B., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., . . . for the, N. I. H. G. E. I. W. p. (2011). Gene-environment interplay in common complex diseases: forging an integrative model—recommendations from an NIH workshop. *Genetic Epidemiology*, n/a-n/a. doi:10.1002/gepi.20571

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., . . . Richardson, S. (2013). GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet, 9*(8), e1003657. doi:10.1371/journal.pgen.1003657

Bottolo, L., & Richardson, S. (2010). Evolutionary Stochastic Search for Bayesian model exploration. *Bayesian Analysis, 5*(3), 508–610.

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology*, n/a–n/a. doi:10.1002/gepi.21965

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell, 169*(7), 1177–1186. doi:10.1016/j.cell.2017.05.038

Breslow, N. E., & Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumor prognosis. *Appl Statist, 48*, 457–468.

Burgess, S., Butterworth, A., & Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol, 37*(7), 658–665. doi:10.1002/gepi.21758

Capanu, M., Concannon, P., Haile, R. W., Bernstein, L., Malone, K. E., Lynch, C. F., . . . Begg, C. B. (2011). Assessment of rare BRCA1 and BRCA2 variants of unknown significance using hierarchical modeling. *Genet Epidemiol, 35*(5), 389–397. doi:10.1002/gepi.20587

Capanu, M., Orlow, I., Berwick, M., Hummer, A. J., Thomas, D. C., & Begg, C. B. (2008). The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Stat Med, 27*(11), 1973–1992.

Casey, G., Conti, D., Haile, R., & Duggan, D. (2013). Next generation sequencing and a new era of medicine. *Gut, 62*(6), 920–932. doi:10.1136/gutjnl-2011-301935

Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., . . . Vermeulen, R. C. (2013). Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen, 54*(7), 542–557. doi:10.1002/em.21797

Chatterjee, N., Park, J.-H., Caporaso, N., & Gail, M. H. (2011). Predicting the Future of Genetic Risk Prediction. *Cancer Epidemiology Biomarkers & Prevention, 20*(1), 3–8. doi:10.1158/1055-9965.epi-10-1022

Cho, I., & Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nat Rev Genet, 13*(4), 260–270. doi:10.1038/nrg3182

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., . . . Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science, 339*(6121), 819–823. doi:10.1126/science.1231143

Conneely, K. N., & Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet, 81*(6), 1158–1168.

Conti, D. V., Lewinger, J. P., Swan, G. E., Tyndale, R. F., Benowitz, N. L., & Thomas, P. D. (2009). Using ontologies in hierarchical modeling of genes and exposures in biologic pathways. In G. E. Swan (Ed.), *Phenotypes and Endophenotypes: Foundations for Genetic Studies of Nicotine Use and Dependence* (Vol. 20, pp. 539–584). Bethesda, MD: NCI Tobacco Control Monographs.

Conti, D. V., & Witte, J. S. (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet, 72*(2), 351–363.

Cordovado, S. K., Hendrix, M., Greene, C. N., Mochal, S., Earley, M. C., Farrell, P. M., . . . Mueller, P. W. (2012). CFTR mutation analysis and haplotype associations in CF patients. *Mol Genet Metab, 105*(2), 249–254. doi:10.1016/j.ymgme.2011.10.013

Cornelis, M. C., Tchetgen Tchetgen, E. J., Liang, L., Qi, L., Chatterjee, N., Hu, F. B., & Kraft, P. (2012). Gene–environment interactions in genome–wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol, 175*(3), 191–202. doi:kwr368 [pii]
10.1093/aje/kwr368

Cortessis, V. K., Thomas, D. C., Levine, A. J., Breton, C. V., Mack, T. M., Siegmund, K. D., . . . Laird, P. W. (2012). Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Hum Genet, 131*(10), 1565–1589. doi:10.1007/s00439–012–1189–8

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High–resolution haplotype structure in the human genome. *Nat Genet, 29*(2), 229–232.

Davey Smith, G., & Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol, 33*(1), 30–42.

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics, 55*(4), 997–1004.

Didelez, V., & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res, 16*(4), 309–330.

Dudbridge, F. (2016). Polygenic Epidemiology. *Genetic Epidemiology*, n/a–n/a. doi:10.1002/gepi.21966

Dudbridge, F., Pashayan, N., & Yang, J. (2017). Predictive accuracy of combined genetic and environmental risk scores. *Genet Epidemiol*. doi:10.1002/gepi.22092

Duffy, D. L., Mitchell, C. A., & Martin, N. G. (1998). Genetic and environmental risk factors for asthma: a cotwin–control study. *American Journal of Respiratory and Critical Care Medicine, 157*(3 Pt 1), 840–845.

Edwards, A. O., Ritter, R., 3rd, Abel, K. J., Manning, A., Panhuysen, C., & Farrer, L. A. (2005). Complement factor H polymorphism and age-related macular degeneration. *Science, 308*(5720), 421-424.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inverence: Algorithms, Evidence and Data Science*. Cambridge, UK: Cambridge University Press.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet, 11*(6), 446-450. doi: nrg2809 [pii]

10.1038/nrg2809

Evans, D. M., Marchini, J., Morris, A. P., & Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genet, 2*(9), e157.

Fehringer, G., Kraft, P., Pharoah, P. D., Eeles, R. A., Chatterjee, N., Schumacher, F. R., . . . African Ancestry Prostate Cancer, C. (2016). Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer Research, 76*(17), 5103-5114. doi: 10.1158/0008-5472.CAN-15-2980

Fejerman, L., Stern, M. C., John, E. M., Torres-Mejia, G., Hines, L. M., Wolff, R. K., . . . Slattery, M. L. (2015). Interaction between common breast cancer susceptibility variants, genetic ancestry, and non-genetic risk factors in Hispanic women. *Cancer Epidemiol Biomarkers Prev*. doi: 10.1158/1055-9965.EPI-15-0392

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinburgh, 52*, 399-433.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., . . . Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., . . . Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science, 296*(5576), 2225-2229.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., . . . Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet, 47*(9), 1091-1098. doi: 10.1038/ng.3367

Gargano, L. M., & Hughes, J. M. (2014). Microbial Origins of Chronic Diseases. *Annual Review of Public Health, 35*(1), 65-82. doi: doi: 10.1146/annurev-publhealth-032013-182426

Gene Ontology, C., Blake, J. A., Dolan, M., Drabkin, H., Hill, D. P., Li, N., . . . Westerfield, M. (2013). Gene Ontology annotations and resources. *Nucleic Acids Res, 41*(Database issue), D530–535. doi:10.1093/nar/gks1050

George, E. I., & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika, 87*, 731–747.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *JASA, 88*, 881–889.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., . . . Nussbaum, R. L. (2003). The International HapMap Project. *Nature, 426*(6968), 789–796.

Gill, J., Obley, A. J., & Prasad, V. (2018). Direct–to–Consumer Genetic Testing: The Implications of the US FDAs First Marketing Authorization for BRCA Mutation Testing. *Jama*. doi:10.1001/jama.2018.5330

Greene, C. S., White, B. C., & Moore, J. H. (2008). Ant colony optimization for genome–wide genetic analysis. In M. Dorigo (Ed.), *Ant Colony Optimization and Swarm Intelligence* (Vol. 5217, pp. 37–47). Berlin: Springer.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *Int J Epidemiol, 29*(4), 722–729.

Gronberg, H., Damber, L., & Damber, J. E. (1994). Studies of genetic factors in prostate cancer in a twin population. *J Urol, 152*(5 Pt 1), 1484–1487; discussion 1487–1489.

GTEx_Consortium. (2015). Human genomics. The Genotype–Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science, 348*(6235), 648–660. doi:10.1126/science.1262110

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., . . . Pericak–Vance, M. A. (2005). Complement factor H variant increases the risk of age–related macular degeneration. *Science, 308*, 419–421.

Hamilton, A. S., & Mack, T. M. (2000). Use of twins as mutual proxy respondents in a case–control study of breast cancer: effect of item nonresponse and misclassification. *American Journal of Epidemiology, 152*(11), 1093–1103.

HapMap. (2003). The International HapMap Project. *Nature, 426*(6968), 789–796.

HapMap. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299–1320.

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., & Li, H. (2013). Sherlock: Detecting Gene–Disease Associations by Matching Patterns of Expression QTL and GWAS. *Am J Hum Genet, 92*(5), 667–680. doi:10.1016/j.ajhg.2013.03.022

Heston, L. L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers. *British Journal of Psychiatry, 112*(489), 819–825.

Hopper, J. L. (1989). Modelling sibship environment in the regressive logistic model for familial disease. *Genetic Epidemiology, 6*(1), 235–240.

Hu, Y. J., & Lin, D. Y. (2010). Analysis of untyped SNPs: maximum likelihood and imputation methods. *Genet Epidemiol, 34*(8), 803–815. doi:10.1002/gepi.20527

Huang, B. E., & Lin, D. Y. (2007). Efficient Association Mapping of Quantitative Trait Loci with Selective Genotyping. *Am J Hum Genet, 80*(3), 567–576.

Hung, R. J., Baragatti, M., Thomas, D., McKay, J., Szeszenia-Dabrowska, N., Zaridze, D., . . . Brennan, P. (2007). Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiol Biomarkers Prev, 16*(12), 2736–2744.

Iglesias, A. I., van der Lee, S. J., Bonnemaijer, P. W. M., Hohn, R., Nag, A., Gharahkhani, P., . . . van Duijn, C. M. (2017). Haplotype reference consortium panel: Practical implications of imputations with large reference panels. *Hum Mutat, 38*(8), 1025–1032. doi:10.1002/humu.23247

Ioannidis, J. P., Boffetta, P., Little, J., OBrien, T. R., Uitterlinden, A. G., Vineis, P., . . . Khoury, M. J. (2008). Assessment of cumulative evidence on genetic associations: interim guidelines. *International Journal of Epidemiology, 37*(1), 120–132.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics, 29*(3), 306–309.

Ioannidis, J. P., Thomas, G., & Daly, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics, 10*(5), 318–329.

Ionita-Laza, I., & Laird, N. M. (2010). On the optimal design of genetic variant discovery studies. *Stat Appl Genet and Mol Biol, 9*(1), Art. 33. doi:10.2202/1544-6115.1581

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., . . . Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res, 36*(Database issue), D480–484.

Katan, M. B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet, 1*(8479), 507–508.

Kaye, J. (2008). The regulation of direct-to-consumer genetic tests. *Hum Mol Genet, 17*(R2), R180–183.

Kerem, B., Rommens, J., Buchanan, J., Markiewicz, D., Cox, T., Chakravarti, A., & Buchwald, M. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science, 245*, 1073–1080.

Kinross, J. M., Darzi, A. W., & Nicholson, J. K. (2011). Gut microbiome-host interactions in health and disease. *Genome Med, 3*(3), 14. doi:10.1186/gm228

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*, 385-389.

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

Kraft, P., Zeggini, E., & Ioannidis, J. P. A. (2009). Replication in Genome-Wide Association Studies. *Statistical Science, 24*(4), 561-573. doi: Doi 10.1214/09-Sts290

Kramer, A., Green, J., Pollard, J., Jr., & Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics, 30*(4), 523-530. doi: 10.1093/bioinformatics/btt703

Kristensen, Frigessi, A., & Borensen, A.-L. (2014). Principles of integrative genomics in cancer. *Nat Rev Cancer*, in press.

Laitinen, T., Rasanen, M., Kaprio, J., Koskenvuo, M., & Laitinen, L. A. (1998). Importance of genetic factors in adolescent asthma: a population-based twin-family study. *American Journal of Respiratory and Critical Care Medicine, 157*(4 Pt 1), 1073-1078.

Lander, E. S., & Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci, 84*, 2363-2367.

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science, 265*, 2037-2048.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., . . . Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature, 467*(7317), 832-838. doi: nature09410 [pii]

10.1038/nature09410

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., . . . Lin, X. (2012). Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *American Journal of Human Genetics, 91*(2), 224-237.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13*(4), 762-775. doi: kxs014 [pii]

10.1093/biostatistics/kxs014

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., & Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol, 31*(8), 871-882.

Li, D., & Conti, D. V. (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol, 169*(4), 497-504.

Li, K., Bihan, M., Yooseph, S., & Methe, B. A. (2012). Analyses of the microbial diversity across the human microbiome. *PLoS One, 7*(6), e32118. doi:10.1371/journal.pone.0032118

Li, R., Conti, D. V., Diaz-Sanchez, D., Gilliland, F., & Thomas, D. C. (2013). Joint Analysis for Integrating Two Related Studies of Different Data Types and Different Study Designs Using Hierarchical Modeling Approaches. *Hum Hered, 74*(2), 83-96. doi:10.1159/000345181

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. (2011). Low coverage sequencing: Implications for the design of complex trait association studies. *Genome Res, 21*(6), 940-951. doi:gr.117259.110 [pii] 10.1101/gr.117259.110

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol, 34*(8), 816-834. doi:10.1002/gepi.20533

Liang, W. E., Thomas, D. C., & Conti, D. V. (2012). Analysis and Optimal Design for Association Studies Using Next-Generation Sequencing With Case-Control Pools. *Genet Epidemiol, 36*(8), 870-881. doi:10.1002/gepi.21681

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., . . . Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med, 343*(2), 78-85. doi:10.1056/NEJM200007133430201

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A Significance Test for the Lasso. *Ann Stat, 42*(2), 413-468 with discussion (pp. 469-517) and rejoinder (pp. 518-431). doi:10.1214/13-AOS1175

Luca, D., Ringquist, S., Klei, L., Lee, A. B., Gieger, C., Wichmann, H. E., . . . Trucco, M. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet, 82*(2), 453-463.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., . . . Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res, 45*(D1), D896-D901. doi:10.1093/nar/gkw1133

MacDonald, M., Lin, C., Srinidhi, L., Bates, G., Altherr, M., Whaley, W., & Lehrach, H. (1991). Complex patterns of linkage disequilibrium in the Huntingtons disease region. *Am J Hum Genet, 49*, 723-734.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., . . . Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science, 339*(6121), 823-826. doi:10.1126/science.1232033

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747–753. doi:nature08494 [pii] 10.1038/nature08494

Marchini, J., Donnelly, P., & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet, 37*(4), 413–417.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499–511. doi:nrg2796 [pii] 10.1038/nrg2796

McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. (2006). Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics, 5*(2), 77–88.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 417–473. doi:10.1111/j.1467-9868.2010.00740.x

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., . . . Guigo, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science, 348*(6235), 660–665. doi:10.1126/science.aaa0355

Minelli, C., De Grandi, A., Weichenberger, C. X., Gögele, M., Modenese, M., Attia, J., . . . Thompson, J. R. (2013). Importance of Different Types of Prior Knowledge in Selecting Genome-Wide Findings for Follow-Up. *Genetic Epidemiology, 37*(2), 205–213. doi:10.1002/gepi.21705

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered, 56*(1–3), 73–82.

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., & White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol, 241*(2), 252–261.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., . . . Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology, 13*(9), R79. doi:10.1186/gb-2012-13-9-r79

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., & Ritchie, M. D. (2008). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genet Epidemiol, 32*(4), 325–340.

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., & et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. *Jama, 315*(1), 68–76. doi:10.1001/jama.2015.17703

Mukherjee, B., Ahn, J., Gruber, S. B., & Chatterjee, N. (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol, 175*(3), 177-190. doi:kwr367 [pii]
10.1093/aje/kwr367

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*(3), 685-694.

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C., & Gauderman, W. J. (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genetic Epidemiology, 35*(3), 201-210. doi:10.1002/gepi.20569

Nebert, D. W., Zhang, G., & Vesell, E. S. (2008). From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metab Rev, 40*(2), 187-224.

Neel, J. (1984). Editorial. *Genet Epidemiol, 1*, 5-6.

Neel, J. V., & Schull, W. J. (1954). *Human Heredity*. Chicago: University of Chicago Press.

OSullivan, O., Coakley, M., Lakshminarayanan, B., Conde, S., Claesson, M. J., Cusack, S., . . . Ross, R. P. (2013). Alterations in intestinal microbiota of elderly Irish subjects post-antibiotic therapy. *The Journal of antimicrobial chemotherapy, 68*(1), 214-221. doi:10.1093/jac/dks348

Omenn, G. S. (2009). From human genome research to personalized healthcare. *Issues in Science and Technology, Summer*, pp 55-61.

Ott, J. (1999). *Analysis of human genetic linkage (3rd ed.)* (Third ed.). Baltimore: Johns Hopkins.

Patel, C. J., Bhattacharya, J., & Butte, A. J. (2010). An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One, 5*(5), e10746. doi:10.1371/journal.pone.0010746

Peer, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol, 32*(4), 381-385. doi:10.1002/gepi.20303

Pepper, J. W., & Rosenfeld, S. (2012). The emerging medical ecology of the human gut microbiome. *Trends Ecol Evol, 27*(7), 381-384. doi:10.1016/j.tree.2012.03.002

Pluzhnikov, A., Below, J. E., Konkashbaev, A., Tikhomirov, A., Kistner-Griffin, E., Roe, C. A., . . . Cox, N. J. (2010). Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am J Hum Genet, 87*(1), 123-128. doi:S0002-9297(10)00308-3 [pii]

10.1016/j.ajhg.2010.06.005

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904-909.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet, 38*, 904-909.

Pritchard, J. K., & Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet, 65*(1), 220-228.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., & Donnelly, P. (2000). Association mapping in structured populations. *Am J Hum Genet, 67*, 170-181.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., ODonovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460*(7256), 748-752. doi:nature08185 [pii]

10.1038/nature08185

Quintana, M. A., Berstein, J. L., Thomas, D. C., & Conti, D. V. (2011). Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet Epidemiol, 35*(7), 638-649. doi:10.1002/gepi.20613

Quintana, M. A., & Conti, D. V. (2013). Integrative variable selection via Bayesian model uncertainty. *Statistics in Medicine, 32*(28), 4928-4953. doi:10.1002/sim.5888

Quintana, M. A., Schumacher, F. R., Casey, G., Bernstein, J. L., Li, L., & Conti, D. V. (2012). Incorporating prior biologic information for high-dimensional rare variant association studies. *Hum Hered, 74*(3-4), 184-195. doi:10.1159/000346021

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat Rev Genet, 12*(8), 529-541. doi:nrg3000 [pii]

10.1038/nrg3000

Rao, D. (1984). Editorial comment. *Genet Epidemiol, 1*, 3.

Raychaudhuri, S., Plenge, R. M., Rossin, E. J., Ng, A. C., Purcell, S. M., Sklar, P., . . . Daly, M. J. (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet, 5*(6), e1000534. doi:10.1371/journal.pgen.1000534

Rebbeck, T. R. (2014). Precision prevention of cancer. *Cancer Epidemiol Biomarkers Prev, 23*(12), 2713-2715. doi:10.1158/1055-9965.EPI-14-1058

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., . . . Gordon, J. I. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science, 341*(6150), 1241214. doi:10.1126/science.1241214

Rieder, M. J., Livingston, R. J., Stanaway, I. B., & Nickerson, D. A. (2008). The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies. *Drug Metab Rev, 40*(2), 241-261.

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

Ročková, V., & George, E. I. (2013). EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association, 109*(506), 828-846. doi:10.1080/01621459.2013.869223

Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., . . . Lanfranchi, G. (2003). Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum Mol Genet, 12*(8), 823-836.

Rothman, K. J., Greenland, S., & Walker, A. M. (1980). Concepts of interaction. *Am J Epidemiol, 112*(4), 467-470.

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., & Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics, 58*(1), 163-170.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., . . . Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet, 37*(7), 710-717. doi:ng1589 [pii] 10.1038/ng1589

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L. R., Peyser, P. A., & Lin, X. (2012). SNP Set Association Analysis for Familial Data. *Genetic Epidemiology, 36*(8), 797.810. doi:10.1002/gepi.21676

Schwabe, R. F., & Jobin, C. (2013). The microbiome and cancer. *Nat Rev Cancer, 13*(11), 800-812. doi:10.1038/nrc3610

Siegmund, K. D., & Gauderman, W. J. (2001). Association tests in nuclear families. *Hum Hered, 52*, 66-76.

Spielman, R. S., McGinnis, R. E., & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet, 52*, 506-516.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A, 100*(16), 9440-9445.

Stunnenberg, H., & Hubner, N. (2014). Genomics meets proteomics: identifying the culprits in disease. *Human Genetics, 133*(6), 689–700. doi:10.1007/s00439–013–1376–2

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge–based approach for interpreting genome–wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545–15550.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., . . . Deloukas, P. (2009). A genome–wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet, 5*(3), e1000433.

Thomas, D. (2000). Genetic epidemiology with a capital "E". *Genet Epidemiol, 19*, 289–3000.

Thomas, D. (2010). Gene–environment–wide association studies: emerging approaches. *Nature Reviews Genetics, 11*(4), 259–272. doi:nrg2764 [pii]
10.1038/nrg2764

Thomas, D. C. (2004). *Statistical methods in genetic epidemiology*. Oxford: Oxford University Press.

Thomas, D. C. (2012). Some surprising twists on the road to discovering the contribution of rare variants to complex diseases. *Hum Hered, 74*(3–4), 113–117. doi:10.1159/000347020

Thomas, D. C. (2017). What Does "Precision Medicine" Have to Say About Prevention? *Epidemiology, 28*(4), 479–483. doi:10.1097/EDE.0000000000000667

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P., & Stram, D. O. (2009). Methodological issues in multistage genome–wide association studies. *Statistical Science, 24*(4), 414–429. doi:Doi 10.1214/09–Sts288

Thomas, D. C., & Clayton, D. G. (2004). Betting odds and genetic associations. *J Natl Cancer Inst, 96*(6), 421–423.

Thomas, D. C., & Witte, J. S. (2002). Point: population stratification: a problem for case–control studies of candidate–gene associations? *Cancer Epidemiol Biomarkers Prev, 11*(6), 505–512.

Thomas, D. C., Yang, Z., & Yang, F. (2013). Two–Phase and Family–Based Designs for Next–Generation Sequencing Studies. *Frontiers in Genetics, 4*, Art 276. doi:10.3389/fgene.2013.00276

Thompson, J. R., Gögele, M., Weichenberger, C. X., Modenese, M., Attia, J., Barrett, J. H., . . . Minelli, C. (2013). SNP Prioritization Using a Bayesian Probability of Association. *Genetic Epidemiology, 37*(2), 214–221. doi:10.1002/gepi.21704

Thompson, S. G., & Willeit, P. (2015). UK Biobank comes of age. *Lancet*. doi:10.1016/S0140-6736(15)60578-5

Thun, M. J., Hoover, R. N., & Hunter, D. J. (2012). Bigger, Better, Sooner,ÄìScaling Up for Success. *Cancer Epidemiology Biomarkers & Prevention, 21*(4), 571-575. doi:10.1158/1055-9965.epi-12-0191

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

Umbach, D. M., & Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Stat Med, 16*(15), 1731-1743. doi:10.1002/(SICI)1097-0258(19970815)16:15<1731::AID-SIM595>3.0.CO;2-S [pii]

Ursell, L. K., Haiser, H. J., Van Treuren, W., Garg, N., Reddivari, L., Vanamala, J., . . . Knight, R. (2014). The intestinal metabolome: an intersection between microbiota and host. *Gastroenterology, 146*(6), 1470-1476. doi:10.1053/j.gastro.2014.03.001

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., de Vos, W. M., . . . Keller, J. J. (2013). Duodenal infusion of donor feces for recurrent Clostridium difficile. *The New England journal of medicine, 368*(5), 407-415. doi:10.1056/NEJMoa1205037

Verkasalo, P. K., Kaprio, J., Koskenvuo, M., & Pukkala, E. (1999). Genetic predisposition, environment and cancer incidence: a nationwide twin study in Finland, 1976-1995. *Int J Cancer, 83*(6), 743-749.

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

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004a). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute, 96*(6), 434-442.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004b). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst, 96*(6), 434-442.

Wacholder, S., McLaughlin, J. K., Silverman, D. T., & Mandel, J. S. (1992). Selection of controls in case-control studies. I. Principles. *Am J Epidemiol, 135*(9), 1019-1028.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet, 81*(2), 208-227.

Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat Rev Genet, 11*(12), 843-854. doi:nrg2884 [pii] 10.1038/nrg2884

Weiss, K. M. (1993). *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge: Cambridge University Press.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Whittemore, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J Appl Statist, 34*((1)), 1-9.

Wild, C. P. (2012). The exposome: from concept to utility. *Int J Epidemiol, 41*(1), 24-32. doi:dyr236 [pii]

10.1093/ije/dyr236

Wilson, M. A., Baurley, J. W., Thomas, D. C., & Conti, D. V. (2010). Complex system approaches to genetic analysis Bayesian approaches. *Adv Genet, 72*, 47-71. doi:B978-0-12-380862-2.00003-5 [pii]

10.1016/B978-0-12-380862-2.00003-5

Witte, J. S., Elston, R. C., & Schork, N. J. (1996). Genetic dissection of complex traits. *Nat Genet, 12*(4), 355-356; author reply 357-358.

Witte, J. S., Gauderman, W. J., & Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol, 149*(8), 693-705.

Wright, F. A., Strug, L. J., Doshi, V. K., Commander, C. W., Blackman, S. M., Sun, L., . . . Cutting, G. R. (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. *Nat Genet, 43*(6), 539-546. doi:ng.838 [pii]

10.1038/ng.838

Wright, S. (1928). Appendix. In P. G. Wright (Ed.), *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.

Xu, C., Wu, K., Zhang, J.-G., Shen, H., & Deng, H.-W. (2017). Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. *Genetic Epidemiology, 41*(3), 187-197. doi:10.1002/gepi.22015

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. doi:ng.608 [pii]

10.1038/ng.608

Zaitlen, N., & Kraft, P. (2012). Heritability in the genome-wide association era. *Hum Genet, 131*(10), 1655-1664. doi:10.1007/s00439-012-1199-6

Zhang, J., & Stram, D. O. (2014). The Role of Local Ancestry Adjustment in Association Studies Using Admixed Populations. *Genetic Epidemiology, 38*(6), 502–515. doi:10.1002/gepi.21835

Zhu, X., Li, S., Cooper, R. S., & Elston, R. C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet, 82*(2), 352–365. doi:10.1016/j.ajhg.2007.10.009

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B, 67*(2), 301–320.