March 3, 2018

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 (Neel and 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 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 (Thomas 2000). 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 (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 discussion below, have made these evaluations possible.

Genetic epidemiology research has been critical to the development of novel statistical methods and 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 (Wright et al. 2011). It has become clear that many complex traits are attributable to a large number of genetic variants, each with a small effect on risk (REF). Deciphering the genetic basis of such polygenic (Purcell et al. 2009; Dudbridge 2016)—or even omnigenic (Boyle et al. 2017)—traits may require very large populations and novel analytic approaches (Thun et al. 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 interaction of cigarette smoking with the *NAT2* gene (Ambrosone et al. 1996) and of ionizing radiation with the *ATM* gene (Bernstein et al.

2010). However, a full understanding of how genes and environment impact 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 within families and across populations. In addressing these fundamental concepts below, we emphasize binary disease traits for simplicity (e.g., cancer), but the general principles are equally applicable to continuous quantitative traits (e.g., blood pressure) with appropriate model specification.

Terminology

The *genome* can be defined as the complete collection of an individual's genetic material. It consists of *chromosomes*, which are long DNA strands present in every cell. 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. 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. Both genic (~1-2% of the genome) and non-genic (~98-99% of the genome) regions can have important effects on disease. A *locus* is a position along the chromosome, which can denote 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). 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 multilocus 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 transmitted from each parent's respective gametes (sperm or ova), which form a complete genotype upon fertilization.

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* 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*. While associations among alleles will decay over generations, associations at neighboring loci will persist within populations, a phenomenon known as *linkage disequilibrium* (LD) (Weiss 1993; Ott 1999). LD 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" (i.e., *familial aggregation*). Clustering within families suggests genetic and/or shared environmental risk factors for a disease. A simple evaluation of aggregation entails 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, which divides the risk for an individual with a particular affected relative by 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 identical (*monozygotic*, MZ) twins than 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, 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 et al. 1994). The authors observed 458 occurrences of prostate cancer. Sixteen MZ and only six DZ twin pairs were concordant for prostate cancer, suggesting that genetics may play a 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 et al. 1999). Nevertheless, a subsequent, much larger study of 44,768 of Swedish, Danish, and Finnish twins (Lichtenstein et al. 2000)(Mucci et al. 2016) concluded that environmental factors may have a larger effect on most cancer sites than germline genetics.

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 likelihoods.

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 (Hopper 1989). A challenge of segregation analysis of disease traits (and particularly rare diseases) relative to continuous traits 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 that conditions on family selection. Of course 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 should generally test for *polygenic* models. Here, the disease is assumed to be caused by many genes, which each have small effects, so one can assume that that their aggregate effect is normally distributed in the population with mean zero and unit variance. Then each individual's polygene has an expectation midway between that of the two parents and variance ½. To compare major gene and polygenic models, both are nested within a 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 either affected pairs of relatives or extended pedigrees, and relies on the phenomenon of 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 run a simple significance test, but cannot estimate the recombination fraction.

The heyday of linkage analysis was the last couple 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 and 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 determine 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 (Risch and Merikangas 1996). 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 power for detecting associations (Huang and Lin 2007). Whatever the phenotype, study subjects should be representative of their respective source populations (Wacholder et al. 1992). In a case-control study, this implies that controls should be individuals who, if diseased, would be cases. Controls are also commonly selected to match cases with respect to ethnicity, age, and sex. 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; Thompson and Willeit 2015). The potential bias arising from using 'convenience' controls is tempered by the low measurement error in SNP genotyping, lack of recall bias when studying inherited variants, large sample sizes, stringent criterion for statistical

significance, and rigorous replication of findings. In addition, if using such controls results in population stratification bias—confounding of associations due to case-control differences in genetic ancestry (Thomas and Witte 2002)—the issue can be addressed analytically with genetic information (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Price et al. 2006a; Zhu et al. 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 being tested is not the causal genetic variant, then the test must be interpreted as a joint test of 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 and 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, a form of McNemar's test for matched pairs (here of alleles) known as the *transmission-disequilibrium test* (Spielman et al. 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 et al. 1999). This is because the increased sharing of genetic information among family members can result in substantially lower power for detecting main effects—although increasing power for detecting gene-environment interactions—than studies of unrelated individuals. Moreover, selecting unrelated cases with a positive family history of disease can improve power in association studies (Antoniou and Easton 2003).

Detecting Association

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). Investigators generally use a regression model to test the variant-disease association that adjusts for confounders. An association that is not considered spurious may have two interpretations (Lander and Schork 1994). First, the disease-associated variant might be the susceptibility allele itself. If so, the association is expected to occur in all populations harboring the allele. 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. In the second case, the marker and disease locus are very close to each other. For this reason, association studies can help 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 the 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 has now lowered the cost for millions of variants to under 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 et al. 2001; Gabriel et al. 2002) (HapMap 2003; HapMap 2005; Frazer et al. 2007). 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*. That this association helped localize a gene within a previously known linkage region demonstrated the power of GWAS.

In the early days of GWAS, the high cost of genotyping motivated the use of multi-stage designs (Satagopan et al. 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 (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 around 1.1–1.3.

Very large sample sizes are required to detect 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 statistically significant GWAS SNPs is somewhat limited, as is the area under the curve (AUC) of receiver operating characteristic (ROC) curves for polygenic risk score predictions. For example, even for height — a highly heritable trait that is easily and accurately measured — a 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 no better than chance). Inclusion of SNPs predicted to be discovered in the future only increased the estimates to 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 (Dudbridge et al. 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). "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 a number of projects (e.g., 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 (Li et al. 2010; Marchini and 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 and 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, including genotype call rates and departures from Hardy-Weinberg equilibrium.

Population Stratification

In the analysis of association study data, one must take potential confounding due to population stratification into account. If variant allele frequencies and disease risks vary across ancestral populations, genetic associations can be confounded. Such confounding can lead to substantial inflation of significance tests. 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 (Price et al. 2006b). Alternatives are to adjust for estimates of the proportion of the genome inherited from different ancestries (Pritchard et al. 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 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 and Stram 2014).

Significance Testing

To determine the overall statistical significance of GWAS results, one must address the issue of multiple comparisons arising from evaluating a million or more effectively independent associations (e.g., variants that are not in strong LD) (Pe'er et al. 2008). The simplest approach is to use a Bonferroni correction, in which the conventional alpha level of $p<0.05$ is divided by the number of effectively independent tests performed (e.g., $0.05/1{,}000{,}000=5\text{x}10^{-8}$). Some GWAS also calculate the false discovery rate (FDR) or use permutation testing to assess the strength of associations (Storey and Tibshirani 2003; Thomas and Clayton 2004; Wacholder et al. 2004a). 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 et al. 1996). A case in point is the increase in heritability explained when considering all variants measured by a GWAS array, and not just those reaching a statistical significance threshold (i.e., the "hidden" heritability). Furthermore, an important aspect of GWAS findings is not simply statistical significance but also their replication in independent samples.

Recent Advances in Association Methods

Next Generation Sequencing

GWAS generally assay known genetic variants with minor allele frequency 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 (e.g., <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. 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 and 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 dismiss an observation of only one or two copies of the variant allele as a false positive.

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 and Laird 2010; Li et al. 2011). An alternative is using DNA pooling. Here one could use either in 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 et al. 2017), or directly in a test of association comparing estimated allele frequencies in case and control pools (Liang et al. 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 et al. 2010; Basu and 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 et al. 2012b). Burden approaches are more powerful if all of the rare variants under study have association effects in the same direction. But 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) (Lee et al. 2012a).

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 (Thomas et al. 2013). This contrasts with a "two phase" design for case-control samples of unrelated individuals (Breslow and 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 significant marginal effects. The 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 a significance level of 1×10^{-13}, and interaction tests require much larger sample sizes than main effects even at the same significance level (Marchini et al. 2005).

Power to discover interaction can be enhanced by "case-only" analyses that assume independence of the interacting factors in the source population. For example, a re-analysis of cleft palate data obtained substantially narrower confidence limits for the interaction between smoking and the *TGF*α gene (odds ratio (OR) 5.14, 95% confidence limits (CI) 1.68–15.7 for the case-only analysis compared with OR 6.57, CI 1.72–20.0 for the case-control analysis), equivalent to a 30% reduction in sample size required for the same precision (Umbach and Weinberg 1997). Case-only analyses are, however, biased if 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.

To overcome these potential biases, various staged and hybrid approaches have been introduced (Evans et al. 2006; Kooperberg and Leblanc 2008; Mukherjee and Chatterjee 2008; Li and Conti 2009; Murcray et al. 2011). Although power will still be much lower for interactions than for main effects, these methods generally yield much better power than a simple exhaustive search (Cornelis et al. 2012; Mukherjee et al. 2012). Moreover, for future studies of gene-environment interactions, it is essential that investigators planning new GWAS design their studies to have appropriate environmental measurements and population-based sampling schemes. For example, the NIH "post-GWAS" initiative aimed at synthesizing all available data on five cancer sites, replicating findings, and characterizing genetic risks and their modification by environmental exposures 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 testing on departures from the multiplicative model, but it has also been recognized that the public health concept of "synergy" is better captured by additive interaction (Rothman et al. 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 (Bernstein et al. 2010), its interaction with *BRCA1* and *BRCA2* is not significantly different from multiplicative but may be somewhat greater than additive (Bernstein et al. 2013). Since carriers of these mutations are at considerably increased baseline risk anyway, a conclusion that their risk would not be further enhanced by radiation exposure may not be appropriate.

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 (Wang et al. 2010; Thomas 2012). 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 processes, 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 et al. 2014). Our understanding of pathways is continually curated into various databases in particular biological domains (e.g., genomic, pathway, functional). Other systems, like the GRAIL program (Raychaudhuri et al. 2009), use statistical techniques based on the theory of 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 (Conti and Witte 2003; Hung et al. 2007; Lewinger et al. 2007; Capanu et al. 2008; Conti et al. 2009; Thomas 2010; Wilson et al. 2010; Capanu et al. 2011). 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 suite of associations under consideration. The inclusion of relatively uninformative annotations in the second-level model produce very little loss of power because such variables received little or no weight, whereas including truly informative annotations greatly improves power (Quintana et al. 2012; Quintana and Conti 2013). 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 (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 as a panel of allergic subjects in a multivariate hierarchical model, along with external pathway information from Ingenuity Pathway Analysis (Li et al. 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 test the causality of the observed association between serum cholesterol and cancer by relating both variables to the *ApoE* genotype. Although the test was never performed, the reference appears to be the first in the medical literature to a technique long used in econometrics (Wright 1928) known as "instrumental variables." In the context of genetic

epidemiology, the basic idea is to indirectly test an association between some biomarker and disease by determining whether a genotype strongly related to the biomarker is also related to disease risk. Here, genotype functions as the instrumental variable (Greenland 2000).

This approach, now known as Mendelian randomization (MR), overcomes two well-known problems in directly testing biomarker-disease associations 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 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 marker-biomarker or marker-disease associations can arise in comparisons across populations. MR does require certain key assumptions (Didelez and 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 marker, 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 testing them for 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 et al. 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 et al. 2016) have been developed that only require half of the variants to be valid instrumental variables.

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 has resulted in an 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; 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 et al. 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, personalized medicine is commonly touted as a major potential benefit of pharmacogenomics and GWAS—albeit with some reservations (Nebert et al. 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.

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.

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 (Manolio et al. 2009; Eichler et al. 2010; Zaitlen and 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 et al. 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 (He et al. 2013; GTEx_Consortium 2015; 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 the recently published result based on RNA samples from 175 individuals across 43 tissues as part of the Genotype-Tissue Expression (GTEx) project (GTEx_Consortium 2015). They highlight tissue-specific and shared regulatory eQTLs and subsequently link 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). This includes measuring high-density panels like the Illumina 450K array, making Epigenome-Wide Association Studies feasible (Rakyan et al. 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 and 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 et al. 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 and Blaser 2012; Schwabe and 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 linking tumor promotion in many cancer types to the effects of bacterial microbiota (Cho and Blaser 2012; Schwabe and Jobin 2013; Gargano and Hughes 2014).

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; Pepper and Rosenfeld 2012; O'Sullivan et al. 2013). 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 (Li et al. 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 et al. 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* (Schadt et al. 2005; Kristensen et al. 2014).

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 significance levels and modifications thereof that allow for the

correlation between tests (Conneely and 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 and Tibshirani 2003). Bayesian versions of the FDR concept have also been proposed (Wacholder et al. 2004b; 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 and 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 (Meinshausen and Bühlmann 2010; Lockhart et al. 2014).

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 and McCulloch 1993; George and Foster 2000). Although not originally proposed for the $p>n$ problem, Bayesian methods of variable selection have recently been extended for that purpose (Bottolo and Richardson 2010; Bottolo et al. 2013; Ročková and George 2013). An example in the context of rare variant association modeling is the Bayesian Risk Index (Quintana et al. 2011)—and extensions using Bayes model averaging in the Bayesian Model Uncertainty method (Quintana and 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 through 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 (Romualdi et al. 2003; McKinney et al. 2006; Moore et al. 2006; Greene et al. 2008; Motsinger-Reif et al. 2008). 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 will undoubtedly expand 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 CB, Freudenheim JL, Graham S, Marshall JR, Vena JE, Brasure JR, Michalek AM, Laughlin R, Nemoto T, Gillenwater KA, Shields PG (1996). Cigarette smoking, n-acetyltransferase 2 genetic polymorphisms, and breast cancer risk. JAMA, 276:1494-501.

Antoniou AC, Easton DF (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. Genet Epidemiol, 25:190-202.

Armitage P, Doll R (1954). The age distribution of cancer and a multistage theory of carcinogenesis. Br J Cancer, 8:1-12.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Antolin M, Artiguenave F, Blottiere HM, Almeida M, Brechot C, Cara C, Chervaux C, Cultrone A, Delorme C, Denariaz G, Dervyn R, Foerstner KU, Friss C, van de Guchte M, Guedon E, Haimet F, Huber W, van Hylckama-Vlieg J, Jamet A, Juste C, Kaci G, Knol J, Lakhdari O, Layec S, Le Roux K, Maguin E, Merieux A, Melo Minardi R, M'Rini C, Muller J, Oozeer R, Parkhill J, Renault P, Rescigno M, Sanchez N, Sunagawa S, Torrejon A, Turner K, Vandemeulebrouck G, Varela E, Winogradsky Y, Zeller G, Weissenbach J, Ehrlich SD, Bork P (2011). Enterotypes of the human gut microbiome. Nature, 473:174-80.

Balliu B, Tsonaka R, Boehringer S, Houwing-Duistermaat J (2015). A retrospective likelihood approach for efficient integration of multiple omics factors in case-control association studies. Genet Epidemiol, 39:156-65.

Bansal V, Libiger O, Torkamani A, Schork NJ (2010). Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet, 11:773-85.

Basu S, Pan W (2011). Comparison of statistical tests for disease association with rare variants. Genet Epidemiol, 35:606-19.

Bernstein JL, Haile RW, Stovall M, Boice JD, Jr., Shore RE, Langholz B, Thomas DC, Bernstein L, Lynch CF, Olsen JH, Malone KE, Mellemkjaer L, Borresen-Dale AL, Rosenstein BS, Teraoka SN, Diep AT, Smith SA, Capanu M, Reiner AS, Liang X, Gatti RA, Concannon P (2010). Radiation exposure, the atm gene, and contralateral breast cancer in the women's environmental cancer and radiation epidemiology study. J Natl Cancer Inst, 102:475-83.

Bernstein JL, Thomas DC, Shore RE, Robson M, Boice JD, Stovall M, Andersson M, Bernstein L, Malone KE, Reiner AS, Lynch CF, Capanu M, Smith SA, Tellhed L, Teraoka SN, Begg CB, Olsen JH, Mellemkjaer L, Liang X, Diep AT, Borg A, Concannon P, Haile RW (2013). Contralateral breast cancer after radiotherapy among brca1 and brca2 mutation carriers: A wecare study report. European journal of cancer (Oxford, England : 1990):Epub 4/18/3.

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

Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, Reedy J, Shaughnessy D, Agurs-Collins T, Paltoo D, Atienza A, Bierut L, Kraft P, Fallin MD, Perera F, Turkheimer E, Boardman J, Marazita ML, Rappaport SM, Boerwinkle E,

Suomi SJ, Caporaso NE, Hertz-Picciotto I, Jacobson KC, Lowe WL, Goldman LR, Duggal P, Gunnar MR, Manolio TA, Green ED, Olster DH, Birnbaum LS, for the NIHGEIWp (2011). Gene-environment interplay in common complex diseases: Forging an integrative model—recommendations from an nih workshop. Genet Epidemiol:n/a-n/a.

Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, Newcombe P, Yengo L, Wild PS, Schillert A, Ziegler A, Nielsen SF, Butterworth AS, Ho WK, Castagne R, Munzel T, Tregouet D, Falchi M, Cambien F, Nordestgaard BG, Fumeron F, Tybjaerg-Hansen A, Froguel P, Danesh J, Petretto E, Blankenberg S, Tiret L, Richardson S (2013). Guess-ing polygenic associations with multiple phenotypes using a gpu-based evolutionary stochastic search algorithm. PLoS Genet, 9:e1003657.

Bottolo L, Richardson S (2010). Evolutionary stochastic search for bayesian model exploration. Bayesian Analysis, 5:508-610.

Bowden J, Davey Smith G, Haycock PC, Burgess S (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol:n/a-n/a.

Boyle EA, Li YI, Pritchard JK (2017). An expanded view of complex traits: From polygenic to omnigenic. Cell, 169:1177-86.

Breslow NE, Chatterjee N (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumor prognosis. Appl Statist, 48:457-68.

Burgess S, Butterworth A, Thompson SG (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol, 37:658-65.

Capanu M, Concannon P, Haile RW, Bernstein L, Malone KE, Lynch CF, Liang X, Teraoka SN, Diep AT, Thomas DC, Bernstein JL, Begg CB (2011). Assessment of rare brca1 and brca2 variants of unknown significance using hierarchical modeling. Genet Epidemiol, 35:389-97.

Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, Begg CB (2008). The use of hierarchical models for estimating relative risks of individual genetic variants: An application to a study of melanoma. Stat Med, 27:1973-92.

Casey G, Conti D, Haile R, Duggan D (2013). Next generation sequencing and a new era of medicine. Gut, 62:920-32.

Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, Liquet B, Vermeulen RC (2013). Deciphering the complex: Methodological overview of statistical models to derive omics-based biomarkers. Environ Mol Mutagen, 54:542-57.

Chatterjee N, Park J-H, Caporaso N, Gail MH (2011). Predicting the future of genetic risk prediction. Cancer Epidemiology Biomarkers & Prevention, 20:3-8.

Cho I, Blaser MJ (2012). The human microbiome: At the interface of health and disease. Nat Rev Genet, 13:260-70.

Conneely KN, Boehnke M (2007). So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. Am J Hum Genet, 81:1158-68.

Conti DV, Lewinger JP, Swan GE, Tyndale RF, Benowitz NL, Thomas PD. Using ontologies in hierarchical modeling of genes and exposures in biologic pathways. In: Swan GE (Eds.). *Phenotypes and endophenotypes: Foundations for genetic studies of nicotine use and dependence*. Bethesda, MD, NCI Tobacco Control Monographs. 2009:539-84.

Conti DV, Witte JS (2003). Hierarchical modeling of linkage disequilibrium: Genetic structure and spatial relations. Am J Hum Genet, 72:351-63.

Cordovado SK, Hendrix M, Greene CN, Mochal S, Earley MC, Farrell PM, Kharrazi M, Hannon WH, Mueller PW (2012). Cftr mutation analysis and haplotype associations in cf patients. Mol Genet Metab, 105:249-54.
Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P (2012). Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol, 175:191-202.
Cortessis VK, Thomas DC, Levine AJ, Breton CV, Mack TM, Siegmund KD, Haile RW, Laird PW (2012). Environmental epigenetics: Prospects for studying epigenetic mediation of exposure-response relationships. Hum Genet, 131:1565-89.
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001). High-resolution haplotype structure in the human genome. Nat Genet, 29:229-32.
Davey Smith G, Ebrahim S (2004). Mendelian randomization: Prospects, potentials, and limitations. Int J Epidemiol, 33:30-42.
Devlin B, Roeder K (1999). Genomic control for association studies. Biometrics, 55:997-1004.
Didelez V, Sheehan N (2007). Mendelian randomization as an instrumental variable approach to causal inference. Stat Meth Med Res, 16:309-30.
Dudbridge F (2016). Polygenic epidemiology. Genet Epidemiol:n/a-n/a.
Dudbridge F, Pashayan N, Yang J (2017). Predictive accuracy of combined genetic and environmental risk scores. Genet Epidemiol.
Edwards AO, Ritter R, 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA (2005). Complement factor h polymorphism and age-related macular degeneration. Science, 308:421-4.
Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet, 11:446-50.
Evans DM, Marchini J, Morris AP, Cardon LR (2006). Two-stage two-locus models in genome-wide association. PLoS Genet, 2:e157.
Fejerman L, Stern MC, John EM, Torres-Mejia G, Hines LM, Wolff RK, Baumgartner KB, Giuliano AR, Ziv E, Perez-Stable EJ, Slattery ML (2015). Interaction between common breast cancer susceptibility variants, genetic ancestry, and non-genetic risk factors in hispanic women. Cancer Epidemiol Biomarkers Prev.
Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler

D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007). A second generation human haplotype map of over 3.1 million snps. Nature, 449:851-61.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002). The structure of haplotype blocks in the human genome. Science, 296:2225-9.

Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Nicolae DL, Cox NJ, Im HK (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat Genet, 47:1091-8.

Gargano LM, Hughes JM (2014). Microbial origins of chronic diseases. Annu Rev Publ Health, 35:65-82.

Gene Ontology C, Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, Peddinti D, Pillai L, Carbon S, Dietze H, Ireland A, Lewis SE, Mungall CJ, Gaudet P, Chrisholm RL, Fey P, Kibbe WA, Basu S, Siegele DA, McIntosh BK, Renfro DP, Zweifel AE, Hu JC, Brown NH, Tweedie S, Alam-Faruque Y, Apweiler R, Auchinchloss A, Axelsen K, Bely B, Blatter M, Bonilla C, Bouguerleret L, Boutet E, Breuza L, Bridge A, Chan WM, Chavali G, Coudert E, Dimmer E, Estreicher A, Famiglietti L, Feuermann M, Gos A, Gruaz-Gumowski N, Hieta R, Hinz C, Hulo C, Huntley R, James J, Jungo F, Keller G, Laiho K, Legge D, Lemercier P, Lieberherr D, Magrane M, Martin MJ, Masson P, Mutowo-Muellenet P, O'Donovan C, Pedruzzi I, Pichler K, Poggioli D, Porras Millan P, Poux S, Rivoire C, Roechert B, Sawford T, Schneider M, Stutz A, Sundaram S, Tognolli M, Xenarios I, Foulgar R, Lomax J, Roncaglia P, Khodiyar VK, Lovering RC, Talmud PJ, Chibucos M, Giglio MG, Chang H, Hunter S, McAnulla C, Mitchell A, Sangrador A, Stephan R, Harris MA, Oliver SG, Rutherford K, Wood V, Bahler J, Lock A, Kersey PJ, McDowall DM, Staines DM, Dwinell M, Shimoyama M, Laulederkind S, Hayman T, Wang S, Petri V, Lowry T, D'Eustachio P, Matthews L, Balakrishnan R, Binkley G, Cherry JM, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hitz BC, Hong EL, Karra K, Miyasato SR, Nash RS, Park J, Skrzypek MS, Weng S, Wong ED, Berardini TZ, Huala E, Mi H, Thomas PD, Chan J, Kishore R, Sternberg P, Van Auken K, Howe D, Westerfield M (2013). Gene ontology annotations and resources. Nucleic Acids Res, 41:D530-5.

George EI, Foster DP (2000). Calibration and empirical bayes variable selection. Biometrika, 87:731-47.

George EI, McCulloch RE (1993). Variable selection via gibbs sampling. J Am Stat Assoc, 88:881-9.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, Huang W, Liu B, Shen Y, Tam PK, Tsui LC, Waye MM, Wong JT, Zeng C, Zhang Q, Chee MS, Galver LM, Kruglyak S, Murray SS, Oliphant AR, Montpetit A, Hudson TJ, Chagnon F, Ferretti V, Leboeuf M, Phillips MS, Verner A, Kwok PY, Duan S, Lind DL, Miller RD, Rice JP, Saccone NL, Taillon-Miller P, Xiao M, Nakamura Y, Sekine A, Sorimachi K, Tanaka T, Tanaka Y, Tsunoda T, Yoshino E, Bentley DR, Deloukas P, Hunt S, Powell D, Altshuler D, Gabriel SB, Zhang H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Aniagwu T, Marshall PA, Matthew O, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Stein LD, Cunningham F, Kanani A, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Donnelly P, Marchini J, McVean GA, Myers SR, Cardon LR, Abecasis GR, Morris A, Weir BS, Mullikin JC, Sherry ST, Feolo M, Daly MJ, Schaffner SF, Qiu R, Kent A, Dunston GM, Kato K, Niikawa N, Knoppers BM, Foster MW, Clayton EW, Wang VO, Watkin J, Sodergren E, Weinstock GM, Wilson RK, Fulton LL, Rogers J, Birren BW, Han H, Wang H, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Todani K, Fujita T, Tanaka S, Holden AL, Lai EH, Collins FS, Brooks LD, McEwen JE, Guyer MS, Jordan E, Peterson JL, Spiegel J, Sung LM, Zacharia LF, Kennedy K, Dunn MG, Seabrook R, Shillito M, Skene B, Stewart JG, Valle DL, Jorde LB, Cho MK, Duster T, Jasperse M, Licinio J, Long JC, Ossorio PN, Spallone P, Terry SF, Lander ES, Nickerson DA, Boehnke M, Douglas JA, Hudson RR, Kruglyak L, Nussbaum RL (2003). The international hapmap project. Nature, 426:789-96.

Greene CS, White BC, Moore JH. Ant colony optimization for genome-wide genetic analysis. In: Dorigo M (Eds.). *Ant colony optimization and swarm intelligence*. Berlin, Springer. 2008:37-47.

Greenland S (2000). An introduction to instrumental variables for epidemiologists. Int J Epidemiol, 29:722-9.

Gronberg H, Damber L, Damber JE (1994). Studies of genetic factors in prostate cancer in a twin population. J Urol, 152:1484-7; discussion 7-9.

GTEx_Consortium (2015). Human genomics. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. Science, 348:648-60.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA (2005). Complement factor h variant increases the risk of age-related macular degeneration. Science, 308:419-21.

HapMap (2003). The international hapmap project. Nature, 426:789-96.

HapMap (2005). A haplotype map of the human genome. Nature, 437:1299-320.

He X, Fuller CK, 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, 92:667-80.

Hopper JL (1989). Modelling sibship environment in the regressive logistic model for familial disease. Genet Epidemiol, 6:235-40.

Hu YJ, Lin DY (2010). Analysis of untyped snps: Maximum likelihood and imputation methods. Genet Epidemiol, 34:803-15.
Huang BE, Lin DY (2007). Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet, 80:567-76.
Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, Zaridze D, Lissowska J, Rudnai P, Fabianova E, Mates D, Foretova L, Janout V, Bencko V, Chabrier A, Moullan N, Canzian F, Hall J, Boffetta P, Brennan P (2007). Inherited predisposition of lung cancer: A hierarchical modeling approach to DNA repair and cell cycle control pathways. Cancer Epidemiol Biomarkers Prev, 16:2736-44.
Iglesias AI, van der Lee SJ, Bonnemaijer PWM, Hohn R, Nag A, Gharahkhani P, Khawaja AP, Broer L, International Glaucoma Genetics C, Foster PJ, Hammond CJ, Hysi PG, van Leeuwen EM, MacGregor S, Mackey DA, Mazur J, Nickels S, Uitterlinden AG, Klaver CCW, Amin N, van Duijn CM (2017). Haplotype reference consortium panel: Practical implications of imputations with large reference panels. Hum Mutat, 38:1025-32.
Ionita-Laza I, Laird NM (2010). On the optimal design of genetic variant discovery studies. Stat Appl Genet and Mol Biol, 9:Art. 33.
Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008). Kegg for linking genomes to life and the environment. Nucleic Acids Res, 36:D480-4.
Katan MB (1986). Apolipoprotein e isoforms, serum cholesterol, and cancer. Lancet, 1:507-8.
Kaye J (2008). The regulation of direct-to-consumer genetic tests. Hum Mol Genet, 17:R180-3.
Kerem B, Rommens J, Buchanan J, Markiewicz D, Cox T, Chakravarti A, Buchwald M (1989). Identification of the cystic fibrosis gene: Genetic analysis. Science, 245:1073-80.
Kinross JM, Darzi AW, Nicholson JK (2011). Gut microbiome-host interactions in health and disease. Genome Med, 3:14.
Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005). Complement factor h polymorphism in age-related macular degeneration. Science, 308:385-9.
Kooperberg C, Leblanc M (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. Genet Epidemiol, 32:255-63.
Kramer A, Green J, Pollard J, Jr., Tugendreich S (2014). Causal analysis approaches in ingenuity pathway analysis. Bioinformatics, 30:523-30.
Kristensen, Frigessi A, Borensen A-L (2014). Principles of integrative genomics in cancer. Nat Rev Cancer:in press.
Lander ES, Green P (1987). Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci, 84:2363-7.
Lander ES, Schork NJ (1994). Genetic dissection of complex traits. Science, 265:2037-48.
Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M,

Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH, Jr., Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467:832-8.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet, 91:224-37.

Lee S, Wu MC, Lin X (2012b). Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13:762-75.

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC (2007). Hierarchical bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol, 31:871-82.

Li D, Conti DV (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol, 169:497-504.
Li K, Bihan M, Yooseph S, Methe BA (2012). Analyses of the microbial diversity across the human microbiome. PLoS One, 7:e32118.
Li R, Conti DV, Diaz-Sanchez D, Gilliland F, Thomas DC (2013). Joint analysis for integrating two related studies of different data types and different study designs using hierarchical modeling approaches. Hum Hered, 74:83-96.
Li Y, Sidore C, Kang HM, Boehnke M, Abecasis G (2011). Low coverage sequencing: Implications for the design of complex trait association studies. Genome Res, 21:940-51.
Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010). Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol, 34:816-34.
Liang WE, Thomas DC, Conti DV (2012). Analysis and optimal design for association studies using next-generation sequencing with case-control pools. Genet Epidemiol, 36:870–81.
Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, 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, 343:78-85.
Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014). A significance test for the lasso. Ann Stat, 42:413-68 with discussion (pp. 69-517) and rejoinder (pp. 8-31).
Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M (2008). On the use of general control samples for genome-wide association studies: Genetic matching highlights causal variants. Am J Hum Genet, 82:453-63.
MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic Acids Res, 45:D896-D901.
MacDonald M, Lin C, Srinidhi L, Bates G, Altherr M, Whaley W, Lehrach H (1991). Complex patterns of linkage disequilibrium in the huntington's disease region. Am J Hum Genet, 49:723-34.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009). Finding the missing heritability of complex diseases. Nature, 461:747-53.
Marchini J, Donnelly P, Cardon LR (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet, 37:413-7.
Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet, 11:499-511.
McKinney BA, Reif DM, Ritchie MD, Moore JH (2006). Machine learning for detecting gene-gene interactions: A review. Appl Bioinformatics, 5:77-88.
Meinshausen N, Bühlmann P (2010). Stability selection. J R Stat Soc Ser B, 72:417-73.
Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, Johnson R, Segre AV, Djebali S, Niarchou A, Consortium GT, Wright FA, Lappalainen T, Calvo M, Getz G, Dermitzakis ET, Ardlie

KG, Guigo R (2015). Human genomics. The human transcriptome across tissues and individuals. Science, 348:660-5.
Minelli C, De Grandi A, Weichenberger CX, Gögele M, Modenese M, Attia J, Barrett JH, Boehnke M, Borsani G, Casari G, Fox CS, Freina T, Hicks AA, Marroni F, Parmigiani G, Pastore A, Pattaro C, Pfeufer A, Ruggeri F, Schwienbacher C, Taliun D, Pramstaller PP, Domingues FS, Thompson JR (2013). Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. Genet Epidemiol, 37:205-13.
Moolgavkar S (1980). Multistage models for carcinogenesis. J Natl Cancer Inst, 65:215-6.
Moore JH (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered, 56:73-82.
Moore JH, Gilbert JC, Tsai CT, Chiang FT, Holden T, Barney N, White BC (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. J Theor Biol, 241:252-61.
Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol, 13:R79.
Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD (2008). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genet Epidemiol, 32:325-40.
Mucci LA, Hjelmborg JB, Harris JR, et al. (2016). Familial risk and heritability of cancer among twins in nordic countries. JAMA, 315:68-76.
Mukherjee B, Ahn J, Gruber SB, Chatterjee N (2012). Testing gene-environment interaction in large-scale case-control association studies: Possible choices and comparisons. Am J Epidemiol, 175:177-90.
Mukherjee B, Chatterjee N (2008). Exploiting gene-environment independence for analysis of case-control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics, 64:685-94.
Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. Genet Epidemiol, 35:201-10.
Nebert DW, Zhang G, Vesell ES (2008). From human genetics and genomics to pharmacogenetics and pharmacogenomics: Past lessons, future directions. Drug Metab Rev, 40:187-224.
Neel J (1984). Editorial. Genet Epidemiol, 1:5-6.
Neel JV, Schull WJ (1954). Human heredity. Chicago,University of Chicago Press.
O'Sullivan O, Coakley M, Lakshminarayanan B, Conde S, Claesson MJ, Cusack S, Fitzgerald AP, O'Toole PW, Stanton C, Ross RP (2013). Alterations in intestinal microbiota of elderly irish subjects post-antibiotic therapy. J Antimicrob Chemother, 68:214-21.
Omenn GS (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.). Baltimore,Johns Hopkins.
Patel CJ, Bhattacharya J, Butte AJ (2010). An environment-wide association study (ewas) on type 2 diabetes mellitus. PLoS One, 5:e10746.

Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol, 32:381-5.
Pepper JW, Rosenfeld S (2012). The emerging medical ecology of the human gut microbiome. Trends in ecology & evolution, 27:381-4.
Pluzhnikov A, Below JE, Konkashbaev A, Tikhomirov A, Kistner-Griffin E, Roe CA, Nicolae DL, Cox NJ (2010). Spoiling the whole bunch: Quality control aimed at preserving the integrity of high-throughput genotyping. Am J Hum Genet, 87:123-8.
Preston-Martin S, Pike MC, Ross RK, Jones PA, Henderson BE (1990). Increased cell division as a cause of human cancer. Cancer Res, 50:7415-21.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006a). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 38:904-9.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006b). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 38:904-9.
Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet, 65:220-8.
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000). Association mapping in structured populations. Am J Hum Genet, 67:170-81.
Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460:748-52.
Quintana MA, Berstein JL, Thomas DC, Conti DV (2011). Incorporating model uncertainty in detecting rare variants: The bayesian risk index. Genet Epidemiol, 35:638-49.
Quintana MA, Conti DV (2013). Integrative variable selection via bayesian model uncertainty. Stat Med, 32:4928-53.
Quintana MA, Schumacher FR, Casey G, Bernstein JL, Li L, Conti DV (2012). Incorporating prior biologic information for high-dimensional rare variant association studies. Hum Hered, 74:184-95.
Rakyan VK, Down TA, Balding DJ, Beck S (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet, 12:529-41.
Rao D (1984). Editorial comment. Genet Epidemiol, 1:3.
Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ (2009). Identifying relationships among genomic disease regions: Predicting genes at pathogenic snp associations and rare deletions. PLoS Genet, 5:e1000534.
Rebbeck TR (2014). Precision prevention of cancer. Cancer Epidemiol Biomarkers Prev, 23:2713-5.
Ridaura VK, Faith JJ, Rey FE, Cheng J, Duncan AE, Kau AL, Griffin NW, Lombard V, Henrissat B, Bain JR, Muehlbauer MJ, Ilkayeva O, Semenkovich CF, Funai K, Hayashi DK, Lyle BJ, Martini MC, Ursell LK, Clemente JC, Van Treuren W, Walters WA, Knight R, Newgard CB, Heath AC, Gordon JI (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science, 341:1241214.

Rieder MJ, Livingston RJ, Stanaway IB, Nickerson DA (2008). The environmental genome project: Reference polymorphisms for drug metabolism genes and genome-wide association studies. Drug Metab Rev, 40:241-61.
Risch N, Merikangas K (1996). The future of genetic studies of complex human diseases. Science, 273:1616-7.
Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015). Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet, 16:85-97.
Ročková V, George EI (2013). Emvs: The em approach to bayesian variable selection. J Am Stat Assoc, 109:828-46.
Romualdi C, Campanaro S, Campagna D, Celegato B, Cannata N, Toppo S, Valle G, 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, 12:823-36.
Rothman KJ, Greenland S, Walker AM (1980). Concepts of interaction. Am J Epidemiol, 112:467-70.
Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB (2002). Two-stage designs for gene-disease association studies. Biometrics, 58:163-70.
Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet, 37:710-7.
Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, Lin X (2012). Snp set association analysis for familial data. Genet Epidemiol, 36:797.810.
Schwabe RF, Jobin C (2013). The microbiome and cancer. Nat Rev Cancer, 13:800-12.
Siegmund KD, Gauderman WJ (2001). Association tests in nuclear families. Hum Hered, 52:66-76.
Spielman RS, McGinnis RE, Ewens WJ (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). Am J Hum Genet, 52:506-16.
Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci, 100:9440-5.
Stunnenberg H, Hubner N (2014). Genomics meets proteomics: Identifying the culprits in disease. Hum Genet, 133:689-700.
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci, 102:15545-50.
Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, Whittaker P, Ranganath V, Kumanduri V, McLaren W, Holm L, Lindh J, Rane A, Wadelius M, Deloukas P (2009). A genome-wide association study confirms vkorc1, cyp2c9, and cyp4f2 as principal genetic determinants of warfarin dose. PLoS Genet, 5:e1000433.
Thomas D (2000). Genetic epidemiology with a capital "e". Genet Epidemiol, 19:289-3000.
Thomas D (2010). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Publ Health, 31:21-36.
Thomas DC (2004). Statistical methods in genetic epidemiology. Oxford,Oxford University Press.

Thomas DC (2012). Some surprising twists on the road to discovering the contribution of rare variants to complex diseases. Hum Hered, 74:113-7.
Thomas DC (2017). What does "precision medicine" have to say about prevention? Epidemiology, 28:479-83.
Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO (2009). Methodological issues in multistage genome-wide association studies. Statist Sci, 24:414-29.
Thomas DC, Clayton DG (2004). Betting odds and genetic associations. J Natl Cancer Inst, 96:421-3.
Thomas DC, Witte JS (2002). Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 11:505-12.
Thomas DC, Yang Z, Yang F (2013). Two-phase and family-based designs for next-generation sequencing studies. Frontiers in Genetics, 4:Art 276.
Thompson JR, Gögele M, Weichenberger CX, Modenese M, Attia J, Barrett JH, Boehnke M, De Grandi A, Domingues FS, Hicks AA, Marroni F, Pattaro C, Ruggeri F, Borsani G, Casari G, Parmigiani G, Pastore A, Pfeufer A, Schwienbacher C, Taliun D, Consortium C, Fox CS, Pramstaller PP, Minelli C (2013). Snp prioritization using a bayesian probability of association. Genet Epidemiol, 37:214-21.
Thompson SG, Willeit P (2015). Uk biobank comes of age. Lancet.
Thun MJ, Hoover RN, Hunter DJ (2012). Bigger, better, sooner,äìscaling up for success. Cancer Epidemiology Biomarkers & Prevention, 21:571-5.
Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological), 58:267-88.
Tomasetti C, Li L, Vogelstein B (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science, 355:1330-4.
Tomasetti C, Vogelstein B (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science, 347:78-81.
Umbach DM, Weinberg CR (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med, 16:1731-43.
Ursell LK, Haiser HJ, Van Treuren W, Garg N, Reddivari L, Vanamala J, Dorrestein PC, Turnbaugh PJ, Knight R (2014). The intestinal metabolome: An intersection between microbiota and host. Gastroenterology, 146:1470-6.
van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, Visser CE, Kuijper EJ, Bartelsman JF, Tijssen JG, Speelman P, Dijkgraaf MG, Keller JJ (2013). Duodenal infusion of donor feces for recurrent clostridium difficile. The New England journal of medicine, 368:407-15.
Verkasalo PK, Kaprio J, Koskenvuo M, Pukkala E (1999). Genetic predisposition, environment and cancer incidence: A nationwide twin study in finland, 1976-1995. Int J Cancer, 83:743-9.
Vogel W (2000). Genetische epidemiologie oder zur spezifität von subdisziplinen der humangenetik. Med Genet, 4:395-99.
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. J Natl Cancer Inst, 96:434-42.

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, 96:434-42.
Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992). Selection of controls in case-control studies. I. Principles. Am J Epidemiol, 135:1019-28.
Wakefield J (2007). A bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet, 81:208-27.
Wang K, Li M, Hakonarson H (2010). Analysing biological pathways in genome-wide association studies. Nat Rev Genet, 11:843-54.
Weiss KM (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, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic Acids Res, 42:D1001-6.
Whittemore AS (2007). A bayesian false discovery rate for multiple testing. J Appl Statist, 34:1-9.
Wild CP (2012). The exposome: From concept to utility. Int J Epidemiol, 41:24-32.
Wilson MA, Baurley JW, Thomas DC, Conti DV (2010). Complex system approaches to genetic analysis bayesian approaches. Adv Genet, 72:47-71.
Witte JS, Elston RC, Schork NJ (1996). Genetic dissection of complex traits. Nat Genet, 12:355-6; author reply 7-8.
Witte JS, Gauderman WJ, Thomas DC (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. Am J Epidemiol, 149:693-705.
Wright FA, Strug LJ, Doshi VK, Commander CW, Blackman SM, Sun L, Berthiaume Y, Cutler D, Cojocaru A, Collaco JM, Corey M, Dorfman R, Goddard K, Green D, Kent JW, Jr., Lange EM, Lee S, Li W, Luo J, Mayhew GM, Naughton KM, Pace RG, Pare P, Rommens JM, Sandford A, Stonebraker JR, Sun W, Taylor C, Vanscoy LL, Zou F, Blangero J, Zielenski J, O'Neal WK, Drumm ML, Durie PR, Knowles MR, Cutting GR (2011). Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2. Nat Genet, 43:539-46.
Wright S. Appendix. In: Wright PG (Eds.). *The tariff on animal and vegetable oils*. New York, Macmillan. 1928.
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. Genet Epidemiol, 41:187-97.
Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010). Common snps explain a large proportion of the heritability for human height. Nat Genet.
Zaitlen N, Kraft P (2012). Heritability in the genome-wide association era. Hum Genet, 131:1655-64.
Zhang J, Stram DO (2014). The role of local ancestry adjustment in association studies using admixed populations. Genet Epidemiol, 38:502-15.
Zhu X, Li S, Cooper RS, Elston RC (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet, 82:352-65.
Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. J R Stat Soc Ser B, 67:301-20.

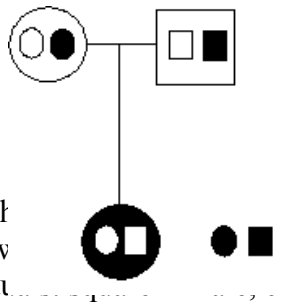

Figure X: Nuclear family with [illegible] lleles transmitted from the parents to the affected child are denoted in v[illegible] mitted from the parents to the affected child are denoted in black. (individu[illegible] cle = female, black = affected, white = unaffected, inside: alleles)