

CIAHD meeting **Thursday March 12, 1:00 - 3:00pm est** (SPH Room 4645)

Agenda:

Continue the stimulating genetics discussion we began at our last meeting. Sharon Kardia will help us understand what genetic epidemiology is doing and together we will brainstorm about how we can meaningfully integrate social and environmental factors into these analyses.

Topics for discussion will include:

1. How can "the environment" be meaningfully incorporated into the candidate gene and GWAS approaches we reviewed last time? What are some challenges in doing this? Can it be done?
2. What is available in Project 2 and are there interesting questions that we could investigate in addition to the ones already proposed? What social or environmental variables could be brought in and how?
3. Are there other data sources we know about that might be of interest to pursue?

***Next Meeting April, 28, 2009, 1:30 – 3:30, Room SPH 4645

Genetic Mapping in Human Disease

David Altshuler,^{1,2,3,4,5*} Mark J. Daly,^{1,2,5*} Eric S. Lander^{1,6,7,8*}

Genetic mapping provides a powerful approach to identify genes and biological processes underlying any trait influenced by inheritance, including human diseases. We discuss the intellectual foundations of genetic mapping of Mendelian and complex traits in humans, examine lessons emerging from linkage analysis of Mendelian diseases and genome-wide association studies of common diseases, and discuss questions and challenges that lie ahead.

By the early 1900s, geneticists understood that Mendel's laws of inheritance underlie the transmission of genes in diploid organisms. They noted that some traits are inherited according to Mendel's ratios, as a result of alterations in single genes, and they developed methods to map the genes responsible. They also recognized that most naturally occurring trait variation, while showing strong correlation among relatives, involves the action of multiple genes and nongenetic factors.

Although it was clear that these insights applied to humans as much as to fruit flies, it took most of the century to turn these concepts into practical tools for discovering genes contributing to human diseases. Starting in the 1980s, the use of naturally occurring DNA variation as markers to trace inheritance in families led to the discovery of thousands of genes for rare Mendelian diseases. Despite great hopes, the approach proved unsuccessful for common forms of human diseases—such as diabetes, heart disease, and cancer—that show complex inheritance in the general population.

Over the past year, a new approach to genetic mapping has yielded the first general progress toward mapping loci that influence susceptibility to common human diseases. Still, most of the genes and mutations underlying these findings remain to be defined, let alone understood, and it remains unclear how much of the heritability of common disease they explain. Below, we discuss the intellectual foundations of genetic mapping, examine emerging lessons, and discuss questions and challenges that lie ahead.

Genetic Mapping by Linkage and Association

Genetic mapping is the localization of genes underlying phenotypes on the basis of correlation with DNA variation, without the need for prior hypotheses about biological function. The simplest form, called linkage analysis, was conceived

by Sturtevant for fruit flies in 1913 (1). Linkage analysis involves crosses between parents that vary at a Mendelian trait and at many polymorphic variants ("markers"); because of meiotic recombination, any marker showing correlated segregation ("linkage") with the trait must lie nearby in the genome.

In the 1970s, the ability to clone and sequence DNA made it possible to tie genetic linkage maps in model organisms to the underlying DNA sequence, and thereby to molecularly clone the genes responsible for any Mendelian trait solely on the basis of their genomic position (2, 3). Such studies typically involved three steps: (i) identifying the locus responsible through a genome-wide search; (ii) sequencing the region in cases and controls to define causal mutation(s); and (iii) studying the molecular and cellular functions of the genes discovered. So-called "positional cloning" became a mainstay of experimental genetics, identifying pathways that are crucial in development and physiology.

Linkage analysis in humans. For most of the 20th century, genome-wide linkage mapping was impractical in humans: Family sizes are small, crosses are not by design, and there were too few classical genetic markers to systematically trace inheritance. Progress in identifying the genes contributing to human traits was initially limited to studies of biological candidates such as blood-type antigens (4) and hemoglobin β protein in sickle-cell anemia (5).

In 1980, Botstein and colleagues, building on their use of DNA polymorphisms to study linkage in yeast (6) and the finding of DNA polymorphism at the globin locus in humans (7, 8), proposed the use of naturally occurring DNA sequence polymorphisms as generic markers to create a human genetic map and systematically trace the transmission of chromosomal regions in families (9). The feasibility of genetic mapping in humans was soon demonstrated with the localization of Huntington disease in 1983 (10). A rudimentary genetic linkage map with ~400 DNA markers was generated by 1987 (11) and was fleshed out to ~5000 markers by 1996 (12). Physical maps providing access to linked chromosomal regions were developed by 1995 (13). With these tools, positional cloning became possible in humans, and the number of disorders tied to a specific gene grew from ~100 in the late 1980s to >2200 today (14).

Several lessons emerged from studies of Mendelian disease genes: (i) The "candidate gene" approach was woefully inadequate; most disease genes were completely unsuspected on the basis

of previous knowledge. (ii) Disease-causing mutations often cause major changes in encoded proteins. (iii) Loci typically harbor many disease-causing alleles, mostly rare in the population. (iv) Mendelian diseases often revealed great complexity, such as locus heterogeneity, incomplete penetrance, and variable expressivity.

Geneticists were eager to apply genetic mapping to common diseases, which also show familial clustering. Mendelian subtypes of common diseases [such as breast cancer (15), hypertension (16), and diabetes (17)] were elucidated, but mutations in these genes explained few cases in the population. In common forms of common disease, risk to relatives is lower than in Mendelian cases, and linkage studies with excellent power to detect a single causal gene yielded equivocal results.

These features were consistent with, but did not prove, a polygenic model. The idea that commonly varying traits might be polygenic in nature was offered by East in 1910 (18). By 1920, linkage mapping was used to identify multiple unlinked factors influencing truncate wings in *Drosophila* (19), and Fisher had developed a mathematical framework for relating Mendelian factors and quantitative traits (20). In the late 1980s, linkage mapping of complex traits was made feasible for experimental organisms through the use of genetic mapping in large crosses (21). But there was little success in humans.

Genetic association in populations. A possible path forward emerged from population genetics and genomics. Instead of mapping disease genes by tracing transmission in families, one might localize them through association studies—that is, comparisons of frequencies of genetic variants among affected and unaffected individuals.

Genetic association studies were not a new idea. In the 1950s, such studies revealed correlations between blood-group antigens and peptic ulcer disease (4); in the 1960s and 1970s, common variation at the human leukocyte antigen (HLA) locus was associated with autoimmune and infectious diseases (22); and in the 1980s, apolipoprotein E was implicated in the etiology of Alzheimer's disease (23). Still, only about a dozen extensively reproduced associations of common variants (outside the HLA locus) were identified in the 20th century (24).

A central problem was that association studies of candidate genes were a shot in the dark: They were limited to specific variants in biological candidate genes, each with a tiny a priori probability of being disease-causing. Moreover, association studies were susceptible to false positives due to population structure, because there was no way to assess differences in the genetic background of cases and controls. Although many claims of associations were published, the statistical support tended to be weak and few were subsequently replicated (25).

In the mid-1990s, a systematic genome-wide approach to association studies was proposed (26–28): to develop a catalog of common human genetic variants and test the variants for associa-

¹Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. ²Center for Human Genetic Research and Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. ³Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA. ⁴Department of Genetics, Harvard Medical School, Boston, MA 02114, USA. ⁵Department of Medicine, Harvard Medical School, Boston, MA 02114, USA. ⁶Department of Systems Biology, Harvard Medical School, Boston, MA 02114, USA. ⁷Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ⁸Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: altshuler@molbio.mgh.harvard.edu (D.A.); mj Daly@chgr.mgh.harvard.edu (M.J.D.); lander@broad.mit.edu (E.S.L.)

tion to disease risk. The focus on common variants as a mapping tool was a matter of practicality, grounded in population genetics. The human population has recently grown exponentially from a small size. As predicted by classical theory (29), humans have limited genetic variation: The heterozygosity rate for single-nucleotide polymorphisms (SNPs) is ~1 in 1000 bases (30–32). Moreover, perhaps 90% of heterozygous sites in each individual are common variants, typically shared among continental populations (33).

If most genetic variation in an individual is common, then why are mutations responsible for Mendelian diseases typically rare? One

answer is natural selection: Mutations that cause strongly deleterious phenotypes—as most Mendelian diseases appear to be—are lost to purifying selection. But if deleterious mutations are typically rare, how could common variants play a role in disease? Common diseases often have late onset, with modest or no obvious impact on reproductive fitness. Mildly deleterious alleles can rise to moderate frequency, particularly in populations that have undergone recent expansion (34). Moreover, some alleles that were advantageous or neutral during human evolution might now confer susceptibility to disease because of changes in living conditions accompanying civi-

lization. Finally, disease-causing alleles could be maintained at high frequency if they were under balancing selection, with disease burden offset by a beneficial phenotype (as in sickle-cell disease and malaria resistance).

These lines of reasoning led to the so-called “common disease–common variant” (CD-CV) hypothesis: the proposal that common polymorphisms (classically defined as having a minor allele frequency of >1%) might contribute to susceptibility to common diseases (26–28). If so, genome-wide association studies (GWASs) of common variants might be used to map loci contributing to common diseases. The concept

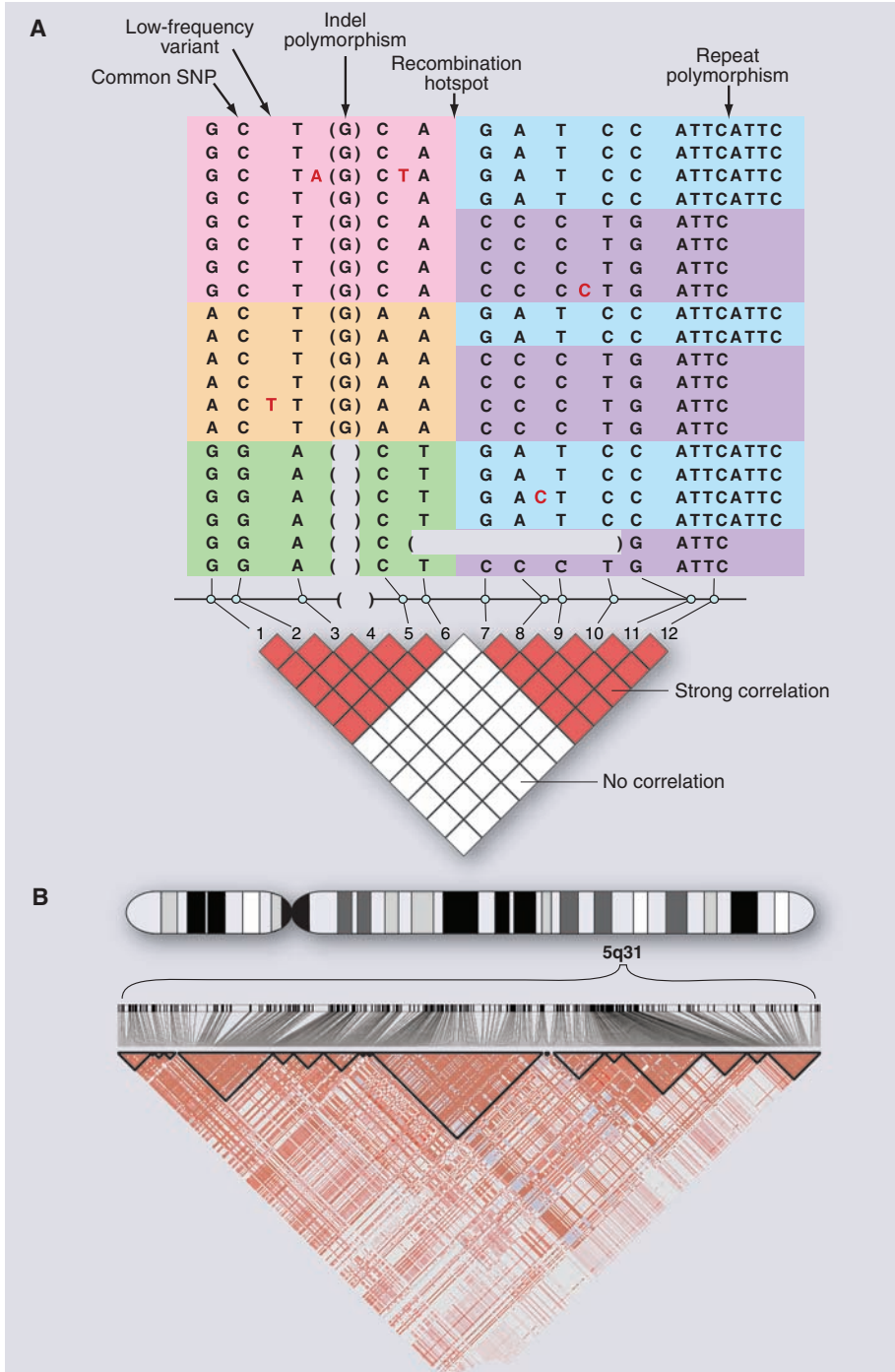


Fig. 1. DNA sequence variation in the human genome. **(A)** Common and rare genetic variation in 10 individuals, carrying 20 distinct copies of the human genome. The amount of variation shown here is typical for a 5-kb stretch of genome and is centered on a strong recombination hotspot. The 12 common variations include 10 SNPs, an insertion-deletion polymorphism (indel), and a tetranucleotide repeat polymorphism on the left side are strongly correlated. Although these six polymorphisms could theoretically occur in 2^6 possible patterns, only three patterns are observed (indicated by pink, orange, and green). These patterns are called haplotypes. Similarly, the six common polymorphisms on the right side are strongly correlated and reside on only two haplotypes (indicated by blue and purple). The haplotypes occur because there has not been much genetic recombination between the sites. By contrast, there is little correlation between the two groups of polymorphisms, because a hotspot of genetic recombination lies between them. The pairwise correlation between the common sites is shown by the red and white boxes below, with red indicating strong correlation and white indicating weak correlation. In addition to the common polymorphisms, lower-frequency polymorphisms also occur in the human genome. Five rare SNPs are shown, with the variant nucleotide marked in red and the reference nucleotide not shown. In addition, on the second to last chromosome, a larger deletion variant is observed that removes several kilobases of DNA. Such larger deletion or duplication events (i.e., CNVs) may be common and segregate as other DNA variants. **(B)** Small regions such as in (A) are often embedded in genomic regions with much greater extents of LD. The diagram shows actual data from the International HapMap Project, showing 420 genetic variants in a region of 500 kb on human chromosome 5q31. Positions of the variants and the pairwise correlations are shown below. Blocks of strong correlation are indicated by the black outlines. Longer-range patterns are often more complex than shown in (A) because weaker recombination hotspots may reduce, but not completely eliminate, marker-to-marker correlation.

Downloaded from www.sciencemag.org on January 29, 2009

was not that all causal mutations at these genes should be common (to the contrary, a full spectrum of alleles is expected), only that some common variants exist and could be used to pinpoint loci for detailed study.

It took a decade to develop the tools and methods required to test the CD-CV hypothesis: (i) catalogs of millions of common variants in the human population, (ii) techniques to genotype these variants in studies with thousands of patients, and (iii) an analytical framework to distinguish true associations from noise and artifacts.

Cataloging SNPs and linkage disequilibrium. Pilot projects in the late 1990s showed that it was possible to identify thousands of SNPs and to perform highly multiplexed genotyping by means of DNA microarrays (35). A public-private partnership, the SNP Consortium, built an initial map of 1.4 million SNPs (32); this has grown to more than 10 million SNPs (36) and is estimated to contain 80% of all SNPs with frequencies of >10% (37).

As the SNP catalog grew, a critical question loomed: Would GWASs require directly testing each of the ~10 million common variants for association to disease? That is, if only 5% of variants were tested, would 95% of associations be missed? Or could a subset serve as reliable proxies for their neighbors? Experience from Mendelian diseases suggested that substantial efficiencies might be possible. Each disease-causing mutation arises on a particular copy of the human genome and bears a specific set of common alleles in cis at nearby loci, termed a haplotype. Because the recombination rate is low [~1 crossover per 100 megabases (Mb) per generation], disease alleles in the population typically show association with nearby marker alleles for many generations, a phenomenon termed linkage disequilibrium (LD) (Fig. 1).

Early studies had demonstrated LD of nearby polymorphisms at the globin locus (38), which proved useful in tracking sickle-cell mutation. In the mid-1980s, it was proposed that a genome-wide search might be performed in genetically isolated populations, scanning the genome for a haplotype shared among unrelated patients carrying the same founder mutation (39). Such "LD mapping" in essence treated the entire population as a very large and very old extended family. This method soon proved useful in fine-mapping the founder $\Delta 508$ mutation in the transmembrane conductance regulator CFTR as a cause of cystic fibrosis (40) and in screening the entire genome in isolated populations such as Finland (41).

The key question was whether the same approach could be used more generally to study common alleles in large human populations, where recombination had more time to whittle down haplotypes. A simulation study suggested that LD might typically be too short to be useful, with a SNP every 5 kb (500,000 SNPs across the genome) providing very weak LD (average correlation $r^2 = 0.1$) (42). Studies of individual loci showed great heterogeneity in local LD (43).

As denser genetic maps became available, a clear picture emerged. Nearby variants were

observed to form a block-like structure consisting of regions characterized by little evidence for historical recombination and limited haplotype diversity (44, 45). Within such regions, which soon proved general (46), genotypes of common SNPs could be inferred from knowledge of only a few empirically determined tag SNPs (45–47). These patterns were shaped by hot and cold spots of recombination in the human genome (48–50), as well as historical population bottlenecks (51).

The International HapMap Project was launched in 2002, with the goal of characterizing SNP frequencies and local LD patterns across the human genome in 270 samples from Europe, Asia, and West Africa. The project genotyped ~1 million SNPs by 2005 (37) and more than 3 million by 2007 (52). Sequence data collected by the project confirmed that the vast majority of common SNPs are strongly correlated to one or more nearby proxies: 500,000 SNPs provide excellent power to test >90% of common SNP variation in out-of-Africa populations, with roughly twice that number required in African populations (37).

Massively parallel genotyping. SNP genotyping was initially performed one SNP at a time, at a cost of ~\$1 per measurement. Multiplex genotyping of hundreds of SNPs on DNA microarrays was demonstrated in 1998 (35), and capacity per array grew from 10,000 to 100,000 SNPs in 2002 to 500,000 to 1 million SNPs in 2007. In parallel, cost fell to \$0.001 per genotype, or less than \$1000 per sample for a whole-genome analysis. By 2006, several technologies could simultaneously genotype hundreds of thousands of SNPs at >99% completeness and >99% accuracy.

Copy-number variation. SNPs are only one type of genetic variation (Fig. 1). Using microarray technology, two groups in 2004 observed that individual copies of the human genome contain large regions (tens to hundreds of kilobases in size) that are deleted, duplicated, or inverted relative to the reference sequence (53, 54). Structural variants had been previously associated with developmental disorders and were often assumed to be pathogenic; the presence of so many segregating copy-number variations (CNVs) in the general population was surprising. The generality of CNVs was soon established (55–59). Many CNVs display tight LD with nearby SNPs (56, 57) and thus can be proxied by nearby SNPs in GWASs. Others occur in regions that are difficult to follow with SNPs, are highly mutable, or are rare (58, 59). Hybrid genotyping platforms have recently been developed to genotype SNPs and CNVs simultaneously (60).

Statistical analysis. Recognizing causal loci amid a genome's worth of random fluctuation required advances in statistical design, analysis, and interpretation. The risk of false negatives was illustrated by a study of type 2 diabetes (T2D) and the Pro¹² → Ala polymorphism in peroxisome proliferator-activated receptor γ . Whereas an initial positive report (61) had not been confirmed in four modest-sized replication studies, larger studies produced strong and consistent evidence of increased risk by a factor of 1.2 (62, 63). The negative studies were actually consistent with

the level of increased risk, but simply lacked adequate power to detect it.

Conversely, stringent thresholds for statistical significance are needed to avoid false positives due to multiple hypothesis testing. Simulations indicated that a dense genome-wide scan of common variants involves the equivalent of ~1 million independent hypotheses (64). A significance level of $P = 5 \times 10^{-8}$ thus represents a finding expected by chance once in 20 GWASs. Large sample sizes would be needed to reach such a stringent threshold (Fig. 2).

Systematic biases could also cause false positives. Differences in ancestry between cases and controls would yield spurious associations (65), suggesting the need for family-based controls (66). It was later recognized that genome-wide studies provide their own internal control: Mismatched ancestry is readily detectable because it produces frequency differences at thousands of SNPs, which could not all reflect causal associations. Methods were developed to detect and adjust for such biases (67–69) as well as unexpected relatedness between subjects. Technical artifacts, which are particularly problematic if cases and controls are not genotyped in parallel (70), were overcome by improved genotyping methods, quality control, and stringent filtering. To maximize efficiency and power, several groups developed methods of selecting tag SNPs (47, 71–73) from empirical LD data and using them to impute genotypes at other SNPs not genotyped in clinical samples (74) on the basis of LD relationships in the HapMap.

Genome-Wide Associations: Lessons

By early 2006, the tools were in place and studies were under way in many laboratories to resolve the hotly debated issue (75, 76) of whether genetic mapping of common SNPs would shed light on common disease. Since then, scores of publications have reported the localization of common SNPs associated with a wide range of common diseases and clinical conditions (age-related macular degeneration, type 1 and type 2 diabetes, obesity, inflammatory bowel disease, prostate cancer, breast cancer, colorectal cancer, rheumatoid arthritis, systemic lupus erythematosus, celiac disease, multiple sclerosis, atrial fibrillation, coronary disease, glaucoma, gallstones, asthma, and restless leg syndrome) as well as various individual traits (height, hair color, eye color, freckles, and HIV viral set point). Figure 3 illustrates data from a paradigmatic genome-wide association study of Crohn's disease performed by the Wellcome Trust Case Control Consortium.

Various lessons have already emerged about genetic mapping by GWAS:

1) GWASs work. Before 2006, only about two dozen reproducible associations outside the HLA locus had been discovered (25). By early 2008, more than 150 relationships were identified between common SNPs and disease traits (table S1). In most diseases studied, GWASs have revealed multiple independent loci, although some traits have not yet yielded associations that meet stringent thresholds (e.g., hypertension). It is not clear whether this

reflects inadequate sample size, phenotypic definition, or a different genetic architecture.

2) Effect sizes for common variants are typically modest. In a few cases, common variants with effects of a factor of ≥ 2 per allele have been found: APOE4 in Alzheimer's disease (23), CFH in age-related macular degeneration (77–79), and LOXL1 in exfoliative glaucoma (80). In the vast majority of cases, however, the estimated effects are much smaller—mostly increases in risk by a factor of 1.1 to 1.5 per associated allele.

3) The power to detect associations has been low. Given the effect sizes now known to exist, and the need to exceed stringent statistical thresholds, the first wave of GWASs provided low power

and height (87–90). Across these four traits and diseases, individual GWASs together documented 29 associations. Increasing the power by pooling the samples to perform meta-analysis and replication genotyping has increased this yield to more than 100 replicated loci for these four conditions.

4) Association signals have identified small regions for study but have not yet identified causal genes and mutations. Genetic mapping is a double-edged sword: Local correlation of genetic variants facilitates the initial identification of a region but makes it difficult to distinguish causal mutation(s). Luckily, whereas family-based linkage methods typically yield regions of 2 to 10 Mb in span, GWASs typically yield more manageable regions of 10 to 100 kb.

already identified seven independent alleles at 8q24 for prostate cancer (92), three at complement factor H (CFH) for age-related macular degeneration (93, 94), three at IRF5 for systemic lupus erythematosus (95), and two at IL23R for Crohn's disease (96). Multiple distinct alleles with different frequencies and risk ratios may well be the rule.

6) A single locus can harbor both common variants of weak effect and rare variants of large effect. In recent GWASs, studies of common SNPs enabled the identification of 19 loci as influencing low- or high-density lipoprotein (LDL, HDL) or triglycerides (84, 85). Nine of these 19 were already known to carry rare Mendelian mutations with large effects, such as the loci for the LDL receptor (LDLR) and familial hypercholesterolemia (FH). Similarly, the genes encoding Kir6.2, WFS1, and TCF2 are all known to cause Mendelian syndromes including T2D, as well as common SNPs with modest effects.

7) Because allele frequencies vary across human populations, the relative roles of common susceptibility genes can vary among ethnic groups. One example is the association of prostate cancer at 8q24: SNPs in the region play a role in all ethnic groups, but the contribution is greater in African Americans. This is not because the risk alleles yet found confer greater susceptibility in African Americans, but because they occur at higher frequencies (92), contributing to the higher incidence among African American men than among men of European ancestry.

Lessons have also emerged about the functions and phenotypic associations of genes related to common diseases:

1) A subset of associations involve genes previously related to the disease. Of 19 loci meeting genome-wide significance in a recent GWAS of LDL, HDL, or triglyceride levels, 12 contained genes with known functions in lipid biology (84, 85). The gene for 3-hydroxy-3-methyl glutaryl-coenzyme A reductase (HMGCR), encoding the rate-limiting enzyme in cholesterol biosynthesis and the target of statin medications, was found by

GWAS to carry common genetic variation influencing LDL levels (84, 85). Similarly, SNPs in the β -cell zinc transporter encoded by SLC30A8 were associated with risk of T2D (97).

2) Most associations do not involve previous candidate genes. In some cases, GWAS results immediately suggest new biological hypotheses—for example, the role of complement factor H in age-related macular degeneration (77–79), FGFR2 in breast cancer (98), and CDKN2A and CDKN2B in T2D (99–101). In many other cases, such as

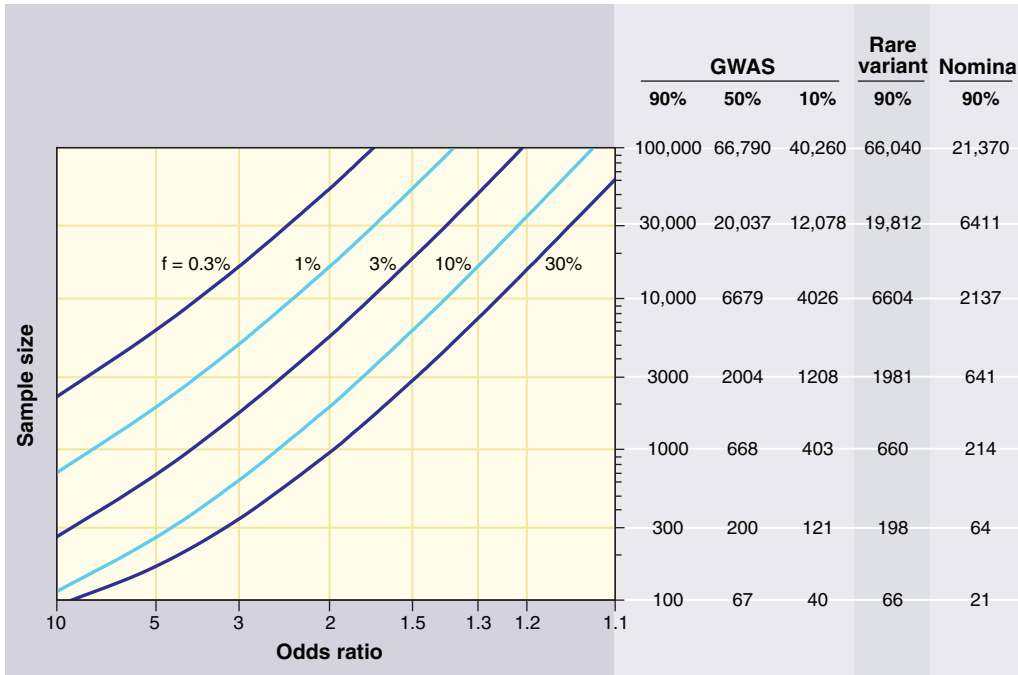


Fig. 2. Sample sizes required for genetic association studies. The graphs show the total number N of samples (consisting of $N/2$ cases and $N/2$ controls) required to map a genetic variant as a function of the increased risk due to the disease-causing allele (x axis) and the frequency of the disease-causing allele (various curves). The required sample size is shown in the table on the right for various different kinds of association studies. The first three columns pertain to GWASs using common variants across the entire genome; the columns correspond to different levels of statistical power to achieve a significant result at $P < 10^{-8}$. The fourth column pertains to a search for rare variants where the frequency listed is the collective frequency of rare variants in controls, and the odds ratio is the excess in cases as compared to controls. Sample sizes assume correction for a genome-wide search of $\sim 20,000$ protein-coding genes in the genome (aiming to achieve $P < 10^{-5}$ with one test performed per gene). The fifth column pertains to a test of a single hypothesis (e.g., testing association with a single SNP). For example, in a GWAS, 1000 samples provide 90% statistical power to detect a 30% allele with a factor of 2 effect. In a genome-wide search via exon sequencing, 660 samples provide 90% power to detect a gene in which rare variants have aggregate population frequency 1% and convey a factor of ~ 8 increase in risk. Note that the sample size to test essentially all common SNPs in the human genome is only 5 times the sample size to test a single SNP.

to discover disease-causing loci (81, 82). For example, achieving 90% power to detect an allele with 20% frequency and a factor of 1.2 effect at a statistical significance of 10^{-8} requires 8600 samples (Fig. 2). Thus, although it is unlikely that common alleles of large effect have been missed, GWASs of hundreds to several thousand cases have necessarily identified only a fraction of the loci that can be found with larger sample sizes. This prediction has been empirically confirmed in T2D (83), serum lipids (84, 85), Crohn's disease (86),

These regions have yet to be scrutinized by fine-mapping and resequencing to identify the specific gene and variants responsible. Even when a locus is identified by SNP association, the causal mutation itself need not be a SNP. For example, the *IRGM* gene was associated with Crohn's disease on the basis of GWAS. Subsequent study suggests that the causal mutation is a deletion upstream of the promoter affecting tissue-specific expression (91).

5) A single locus can contain multiple independent common risk variants. Intensive study has

LOC387715/HTRA1 with age-related macular degeneration (102), nearby genes have no known function.

3) Many associations implicate non–protein-coding regions. Although some associated non-coding SNPs may ultimately prove attributable to LD with nearby coding mutations, many are sufficiently far from nearby exons to make this outcome unlikely. Examples include the region at 8q24 associated with prostate, breast, and colon cancer, 300 kb from the nearest gene (103, 104), and the region at 9q21 associated with myocardial infarction and T2D, 150 kb from the nearest genes encoding CDKN2A and CDKN2B (99–101, 105–107).

A role for noncoding sequence in disease risk is not surprising: Comparative genome analysis has shown that 5% of the human genome is evolutionarily conserved and thus functional; less than one-third of this 5% consists of genes that encode proteins (108). Noncoding mutations with roles in disease susceptibility will likely open new doors to understanding genome biology and gene regulation. Regulatory variation also suggests different therapeutic strategies: Modulating levels of gene expression may prove more tractable than replacing a fully defective protein or turning off a gain-of-function allele.

4) Some regions contain expected associations across diseases and traits. Crohn's disease, psoriasis, and ankylosing spondylitis have long been recognized to share clinical features; the association of the same common polymorphisms in IL23R in all three diseases points to a shared molecular cause (96, 109, 110). SNPs in STAT4 (signal transducer and activator of transcription 4) are associated with rheumatoid arthritis and systemic lupus, two diseases that share clinical features. Multiple variants associated with T2D are associated with insulin secretion defects in nondiabetic individuals (101, 111–116), highlighting the role of β -cell failure in the pathogenesis of T2D.

5) Some regions reveal surprising associations. For example, unexpected connections have emerged among T2D, inflammatory diseases (two loci), and cancer (four loci). A single intron of CDKAL1 was found to contain a SNP associated with T2D and insulin secretion defects (99–101, 116), and another with Crohn's disease and psoriasis (117). A coding variant in glucokinase regulatory protein is associated with triglyceride levels and fasting glucose (101) but also with C reactive protein levels (118, 119) and Crohn's disease (86). A SNP in TCF2 is associated with protection from T2D, as expected on the basis of Mendelian mutations at the same gene (120). Unexpectedly, the same association signal turned up in a GWAS for prostate cancer (121). Similarly, JAZF1 was identified as containing SNPs associated with T2D (83) and prostate cancer (122), and TCF7L2 with T2D (123) and colon cancer (124, 125).

From Common SNPs to the Full Allelic Spectrum

The current HapMap provides reliable proxies for the vast majority of SNPs at frequencies above

5%, but its coverage declines rapidly for lower-frequency alleles (37). Such lower-frequency alleles may be particularly important: Alleles with strong deleterious effects are constrained by natural selection from becoming too common. We divide these alleles into two conceptually distinct classes:

1) Common variants with frequencies below 5%. By “common,” we refer to variants that occur at sufficient frequency to be cataloged in studies of the general population and measured (directly, or indirectly through LD) in association studies. In practice, this class may include allele frequencies in the range of 0.5% and above. A GWAS of 2000 cases and 2000 controls provides good power for a 1% allele causing a factor of 4 increase in risk (even at $P < 10^{-8}$) (Fig. 2).

The value of lower-frequency common variants is illustrated by PCSK9 (proprotein convertase subtilisin/kexin type 9). The gene encoding PCSK9 contains very rare mutations causing autosomal dominant hypercholesterolemia (discovered by linkage analysis), as well as high-frequency common variants with modest effects. The former are too rare and the latter too weak to enable effective clinical study of PCSK9 with respect to coronary artery disease risk. Hobbs and Cohen sequenced the gene (126, 127) and identified low-frequency common variants (0.5 to 1%), which allowed epidemiological research documenting a protective effect on myocardial infarction (128).

2) Rare variants. Most Mendelian diseases involve rare mutations that are essentially never observed in the general population. Rare mutations likely also play an important role in common diseases. Because they are numerous and individually rare, it is not possible to create a complete catalog in the general population. Instead, they must be identified by sequencing in cases and controls in each study. Moreover, because each variant is too rare to prove statistical evidence of association, the mutations must be aggregated as a class to compare the overall frequency of cases versus controls.

A few examples are known through candidate gene studies. Rare nonsynonymous mutations in MC4R are found in patients with extreme early-onset obesity (129). Rare nonsynonymous mutations in ABCA1 are more common in patients with extremely low HDL than in those with high HDL (130). An excess of rare mutations in renal salt-handling genes has been associated with lower blood pressure and protection against hypertension (131).

The sample size required to perform a genome-wide search based on coding mutations depends on the background frequency (μ) of mutations that confer disease risk and the level (ω) of increased risk for each such mutation. ABCA1 is a favorable case because μ and ω are high (the gene has an unusually large coding region of ~ 7 kb, and mutations confer a factor of ~ 6 increase in risk). Achieving genome-wide significance will likely require resequencing studies of thousands of cases and controls, similar to GWASs (Fig. 2).

GWASs of rare variants are already under way for large structural variants through the use of microarray analysis. A recent GWAS of autism revealed

that a highly penetrant, recurrent microdeletion and microduplication of a 593-kb region in 16p11.2 explains 1% of cases (132). Moreover, several recent studies report that patients with autism and schizophrenia may have an excess of rare deletions across the genome relative to unaffected controls (133, 134). Although these studies did not identify specific loci (none of the novel loci were observed more than once), they suggest that the universe of rare structural changes contributing to each disease may be as large and diverse as that of common SNPs.

The Genetic Architecture of Common Disease

Variants so far identified by GWASs together explain only a small fraction of the overall inherited risk of each disease (for example, $\sim 10\%$ of the variance for Crohn's and $\sim 5\%$ for T2D). Where is the remaining genetic variance to be found? There are several answers:

1) At disease loci already identified by GWAS, the locus-attributable risk will often be higher than currently estimated. This is because marker SNPs used in GWASs will typically be imperfect proxies for the actual causal mutation that led to the association signal. The causal gene will often contain additional mutations not tagged by the initial marker SNPs, both common and rare. Determining the contribution of each gene will require intensive studies of variants at each locus.

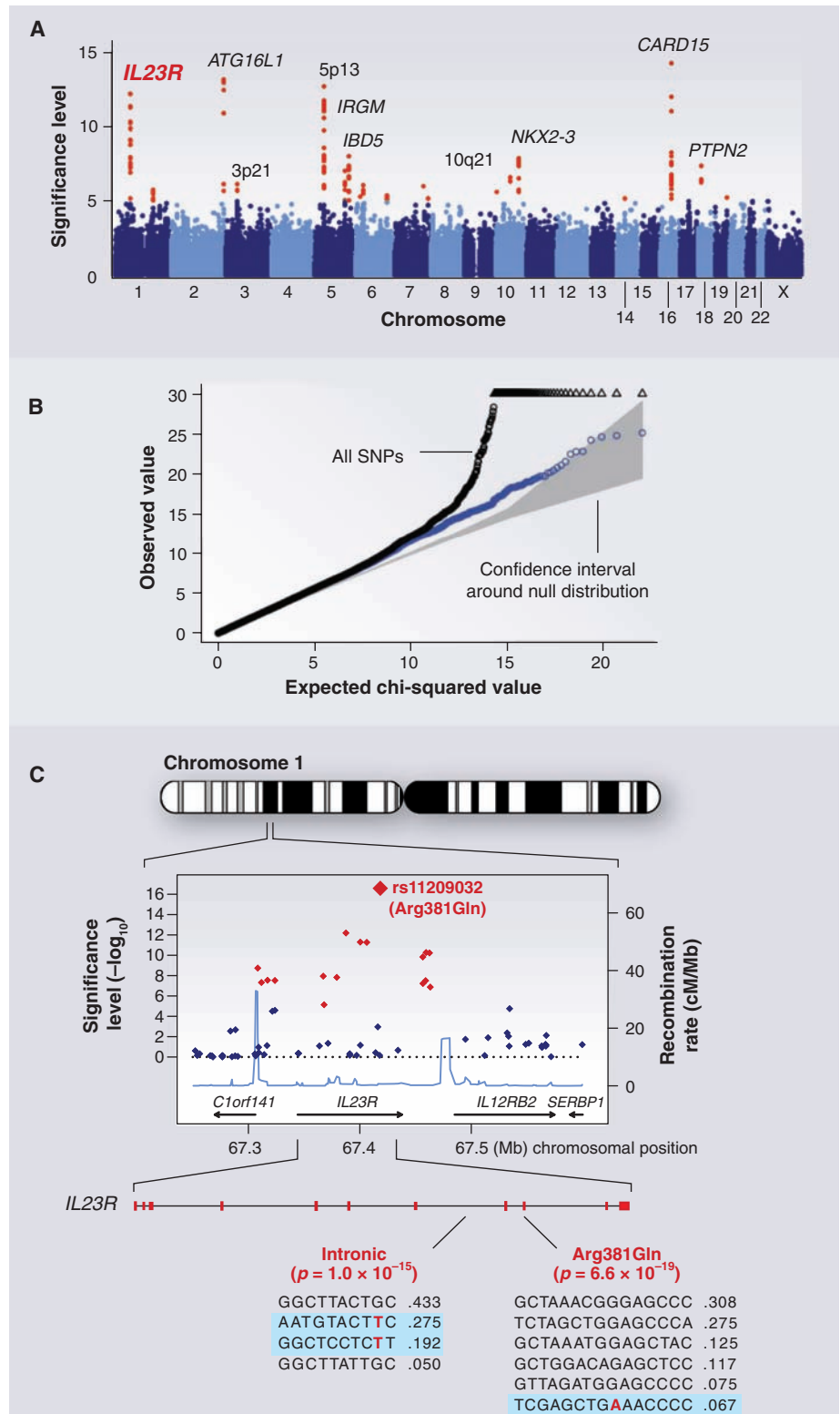
2) Many more disease loci remain to be identified by GWAS. As noted above, GWASs to date have had low statistical power and thus necessarily missed many loci with common variants of similar and smaller effects. The first studies did not have proxies for common structural variants and have failed to capture lower-frequency common variants (0.5 to 5%). Moreover, the vast majority of studies have been performed only in samples of European ancestry. Larger, more comprehensive, and more diverse GWASs will reveal many more loci.

3) Some disease loci will contain only rare variants. Such loci (if not already found by Mendelian genetics) cannot be identified by study of common variants alone. They will require systematic resequencing of all genes in large samples (Fig. 2).

4) Current estimates of the variance explained are based on simplifying assumptions. Because the genotype-phenotype correlation has yet to be well characterized, the estimates assume that the variants interact in a simple additive manner. Yet gene-gene and gene-environment interactions play important roles in disease risk. Although searches have not yet found much evidence for epistasis [e.g., (93, 94, 135)], this may simply reflect limited power to assess the many possible modes of interaction, including pairwise interactions and threshold effects. Once patterns of association and interaction are understood, effects of specific gene and environmental exposures on each phenotype may be larger.

For these reasons, it is premature to make inferences about the overall genetic architecture of common disease. Only by systematically exploring each of these directions over the coming years will a general picture emerge—with the likely

Fig. 3. GWAS for Crohn's disease. The panels show data from the study of Crohn's disease by the Wellcome Trust Case Control Consortium. **(A)** Significance level (P value on \log_{10} scale) for each of the 500,000 SNPs tested across the genome. SNP locations reflect their positions across the 23 human chromosomes. SNPs with significance levels exceeding 10^{-5} (corresponding to 5 on the y axis) are colored red; the remaining SNPs are in blue. Ten regions with multiple significant SNPs are shown, labeled by their location or by the likely disease-related gene (e.g., *IL23R* on chromosome 1). **(B)** The fact that the SNPs in red are extreme outliers is made clear from a so-called Q-Q plot. A Q-Q plot is made as follows: The SNPs are ordered (from 1 to n) according to their observed P values; observed and expected P values are plotted for each SNP. Under the null distribution, the expected P value for the i th SNP is i/n . If there are no significant associations, the Q-Q plot will lie along the 45° line; the gray region corresponds to a 95% confidence region around this null expectation. Black points correspond to all 500,000 SNPs studied that passed strict quality control; they diverge strongly from the null expectation. Blue points reflect the P values that remain when the SNPs in the 10 most significant regions are removed; there is still some excess of significant P values, indicating the presence of additional loci of more modest effect. **(C)** Close-up of the region around the *IL23R* locus on chromosome 1. The first part shows the significance levels for SNPs in a region of ~ 400 kb, with colors as in (A). The highest significance level occurs at a SNP in the coding region of the *IL23R* gene (causing an Arg³⁸¹ \rightarrow Gln change). The light blue curve shows the inferred local rate of recombination across the region. There are two clear hotspots of recombination, with SNPs lying between these hotspots being strongly correlated in a few haplotypes. The second part shows that the *IL23R* locus harbors at least two independent, highly significant disease-associated alleles. The first site is the Arg³⁸¹ \rightarrow Gln polymorphism, which has a single disease-associated haplotype (shaded in blue) with frequency of 6.7%. The second site is in the intron between exons 7 and 8; it tags two disease-associated haplotypes with frequencies of 27.5% and 19.2%.



outcome being that different diseases will each be characterized by a different balance of allele frequencies, interactions, and types. Although the proportion of genetic variance explained is certain to grow in the coming years, it is unlikely to approach 100% because of practical limitations, such as the difficulty of detecting common variants with extremely small effects, genes harboring rare var-

iants at very low frequency, and complex interactions among genes and with the environment.

Disease Risk Versus Disease Mechanism

The primary value of genetic mapping is not risk prediction, but providing novel insights about mechanisms of disease. Knowledge of disease pathways (not limited to the causal genes and mutations) can

suggest strategies for prevention, diagnosis, and therapy. From this perspective, the frequency of a genetic variant is not related to the magnitude of its effect, nor to the potential clinical value that may be obtained.

The classic example is Brown and Goldstein's studies of FH, which affects $\sim 0.2\%$ of the population and accounts for a tiny fraction of the heritability of LDL and myocardial infarction. Studies of FH led

to the discovery of the LDL receptor and supported the development of HMGCR inhibitors (statins) for lowering LDL, the use of which is not limited to FH carriers.

More recently, GWASs have shown that common genetic variation in LDLR and HMGCR influences LDL levels (84, 85). Although SNPs in HMGCR have only a small effect (~5%) on LDL levels, drugs targeting the encoded protein decrease LDL levels by a much greater extent (~30%). This is because the effect of an inherited variant is limited by natural selection and pleiotropy, whereas the effect of a drug treatment is not.

The Path Ahead

Given the long-standing success of genetic mapping in providing new insights into biology and disease etiology, and the recent proof that systematic association studies can identify novel loci, our aim should be nothing less than identifying all pathways at which genetic variation contributes to common diseases. We sketch key steps in achieving this goal.

Expanding clinical studies. Current studies are underpowered for the types of SNP alleles that we now know exist, and available evidence indicates that increasing sample size will yield substantial returns. A study of 1000 cases and 1000 controls provides only 1% power to detect a 20% variant that increases risk by a factor of 1.3, but a study of 5000 cases and 5000 controls provides 98% power (Fig. 2). Moreover, early data on rare single-nucleotide (130, 131) and structural variants (133, 134, 136) indicate that similarly large samples will be needed to achieve the levels of statistical significance required to detect rare events in a genome-wide search.

Nearly all GWASs to date have been performed in populations of European ancestry. Even if a variant has the same effect in all ancestry groups, it may be more readily detected in one population simply because it happens to have higher frequency. Genetic effects will likely vary across groups because of modification by environment and behavior, which may vary more across groups than does genotype.

Many important diseases remain to be studied by GWAS. Disease-related intermediate traits can also offer substantial insight, particularly in conjunction with clinical endpoints. For example, newly described variants on chromosome 1 (near SORT1) are associated both with levels of LDL cholesterol (84, 85) and with risk of myocardial infarction (106); this provides not only increased statistical confidence, but also a biomarker for gene function and pathophysiological insight. Genetic variants that influence gene expression [e.g., (137)] hold promise for elucidating regulatory pathways. Mapping of modifiers of Mendelian mutations—for example, genes that influence the age of onset in carriers of BRCA1 and BRCA2 mutations—may suggest ways to reverse high risk due to mutations.

Correlations between genetic variants and phenotypes are limited by the accuracy with which

each is measured. The ability to measure genotype now far exceeds our ability to measure phenotype. Continuous ambulatory monitoring, imaging methods, and comprehensive (“-omic”) approaches to biological samples all have promise in improving the accuracy of phenotype measurement.

Environmental exposures play a larger role in human phenotypic variation than does genetic variation, but environmental exposures are fundamentally more difficult to measure. DNA is stable throughout life, with a single physical chemistry that enables generic approaches to measurement. Environmental exposures are heterogeneous and may be fleeting. Improved methods for measuring environmental exposures, perhaps based on epigenetic marks they leave, are sorely needed.

Expanding the range of genetic variation. The lowest-hanging fruit will be to resequence loci that have been definitively implicated in disease by Mendelian genetics or by GWAS. Because the prior probability of a true association is higher, such regions will be the best setting to develop methods for understanding the statistical significance and biological importance of rare mutations. Initially, resequencing of coding exons will be easiest to interpret. Rare coding mutations with large effect will be especially valuable, because physiological studies of mutation carriers can help illuminate the biological basis of the disease, and because coding mutations of large effect are more straightforwardly transferred to cellular and animal models for mechanistic studies.

Extending GWASs to include structural variants and lower-frequency common variants will require comprehensive catalogs of genomic variation, as well as characterization of LD relationships. With new massively parallel sequencing technologies, an accurate map of all 1% alleles (both single-nucleotide and structural) should be achievable. A “1000 Genomes Project” was recently launched toward this end (138).

Some loci may harbor neither common variants nor rare structural variants, and thus will be missed by array and LD-based approaches. Discovering such genes will require sequencing in thousands of cases and controls. Initial studies will likely focus on exons, where functional mutations are enriched to the greatest extent. Highly parallel methods to capture hundreds of thousands of exons, and other targets of interest, are under development (139).

Multiple instances of de novo coding mutations at a locus (by comparing affected individuals with parents) could provide particularly powerful association information, because the human mutation rate is so low (in the range of 10^{-8}). But identifying de novo mutations without being overwhelmed by false positives will require extraordinary sequencing accuracy (far better than finished genome sequence). Because such studies will be expensive at first, priority should go to disorders with high heritability, where there is an unmet medical need, and for which other approaches

have met with limited success. Psychiatric disorders might represent one such target.

Eventually, it will become practical to resequence entire genomes from thousands of cases and controls. The problem of interpretation will be much harder for noncoding functional elements, because it is unclear either how to aggregate elements to achieve a large enough target size, or to develop ways to recognize function-altering changes.

Routine genome sequencing of deeply phenotyped cohorts will fundamentally change the nature of genetic mapping: from the current serial process (in which initial localization by linkage or GWAS is followed by scrutiny of DNA variation and phenotypes) to a joint estimation procedure combining variation information of all types, frequencies, and phenotypes to discover and characterize genotype-phenotype correlations. New statistical methods will be required to combine evidence from rare and common alleles at a locus and across multiple loci, phenotypes, and non-genetic exposures. A particular challenge will be to identify mutations in regions without known function or evolutionary conservation.

There may be inherent limits to our ability to relate phenotypic variation and genotypic variation. To the extent that disease is influenced by tiny effects at hundreds of loci or highly heterogeneous rare mutations, it may be impractical to assemble sufficiently large samples to give a complete accounting.

Implications for Biology, Medicine, and Society

Genetic mapping is only a first step toward biological understanding and clinical application. Useful tools will include maps of evolutionary conservation (108) and chromatin state (140), as well as databases of cell-state signatures, such as genome-wide expression patterns, that may integrate aspects of cell biology under resting and provoked conditions (141). Creation of disease models, both in human cell culture and nonhuman animals, will be key. Physiological studies in patients classified by genotype may inform disease processes and lead to useful nongenetic biomarkers. Given the limits of human clinical research, rare alleles of strong effect may be more useful than common alleles of weak effect.

The high failure rate of clinical trials testifies to the limited predictive value of current approaches. By focusing attention on genes and processes, human genetics has the potential to yield productive targets and predictive animal models. In clinical trials, the ability to stratify patients by genotype or biological pathway may reveal differences in therapeutic response. Genetics may also increase the efficiency of outcome trials by focusing on patients at higher-than-average risk.

The extent to which genetic information will figure in “personalized medicine” will depend on whether predictive accuracy beyond conventional measures can be attained, and whether there are interventions whose effectiveness is improved by knowledge of a genetic test. Knowledge of a common variant that increases T2D

risk by 20% may eventually lead to new understanding and therapeutic strategies, but whether an increase in absolute risk (from 8% to 10%) is useful for patients remains to be seen. Although it is tempting to think that knowledge of individual risk might promote greater adherence to a healthy lifestyle, human behavior is complex and risk estimates are challenging to interpret. Even where genotype can predict response to a drug with a narrow therapeutic window, it cannot be assumed that genetic testing will necessarily lead to improved clinical outcomes.

Our understanding of complex disease will be in constant flux over the coming years. The pace of discovery, while scientifically exhilarating, poses daunting challenges. Direct-to-consumer marketing of genetic information is already under way. It will be a challenge for the public to understand the difference between relative and absolute risk, and to figure in their thinking the larger component of genetic and environmental factors not yet captured by today's technologies. Rigorous assessment of health benefit and cost are needed, including costs of testing and treatment that may flow from an altered sense of risk. As genetic information is shown to be useful, equitable access will be critical.

Finally, we must ensure that the promise of research on genetic factors in complex disease does not encourage a mistaken sense of genetic determinism. This is especially important for behavioral traits, which are especially prone to misinterpretation and misguided policy. We must constantly remind the public—and ourselves—that although genes play a role (and can lead us to new biological insight), our traits are powerfully shaped by the environment, and the solutions to important problems will often lie outside our genes.

References

1. A. Sturtevant, *J. Exp. Zool.* **14**, 43 (1913).
2. L. Clarke, J. Carbon, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2173 (1980).
3. W. Bender *et al.*, *Science* **221**, 23 (1983).
4. I. Aird, H. H. Bental, J. A. Mehigan, J. A. Roberts, *Br. Med. J.* **2**, 315 (1954).
5. V. M. Ingram, *Nature* **178**, 792 (1956).
6. T. D. Petes, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5091 (1977).
7. A. J. Jeffreys, *Cell* **18**, 1 (1979).
8. Y. W. Kan, A. M. Dozy, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 5631 (1978).
9. D. Botstein, R. L. White, M. Skolnick, R. W. Davis, *Am. J. Hum. Genet.* **32**, 314 (1980).
10. J. F. Gusella *et al.*, *Nature* **306**, 234 (1983).
11. H. Donis-Keller *et al.*, *Cell* **51**, 319 (1987).
12. C. Dib *et al.*, *Nature* **380**, 152 (1996).
13. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
14. Online Mendelian Inheritance in Man (www.ncbi.nlm.nih.gov/sites/entrez?db=omim).
15. P. L. Welch, M. C. King, *Hum. Mol. Genet.* **10**, 705 (2001).
16. R. P. Lifton, *Harvey Lect.* **100**, 71 (2004).
17. G. I. Bell, K. S. Polonsky, *Nature* **414**, 788 (2001).
18. E. East, *Am. Nat.* **44**, 65 (1910).
19. E. Altenburg, H. J. Muller, *Genetics* **5**, 1 (1920).
20. R. A. Fisher, *Trans. R. Soc. Edinburgh* **52**, 399 (1918).
21. A. H. Paterson *et al.*, *Nature* **335**, 721 (1988).
22. J. Klein, A. Sato, *N. Engl. J. Med.* **343**, 782 (2000).
23. W. J. Strittmatter, A. D. Roses, *Annu. Rev. Neurosci.* **19**, 53 (1996).
24. J. N. Hirschhorn, K. Lohmueller, E. Byrne, K. Hirschhorn, *Genet. Med.* **4**, 45 (2002).
25. K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, J. N. Hirschhorn, *Nat. Genet.* **33**, 177 (2003).
26. F. S. Collins, M. S. Guyer, A. Chakravarti, *Science* **278**, 1580 (1997).
27. E. S. Lander, *Science* **274**, 536 (1996).
28. N. Risch, K. Merikangas, *Science* **273**, 1516 (1996).
29. M. Kimura, T. Ota, *Genetics* **75**, 199 (1973).
30. H. Harris, *Proc. R. Soc. London Ser. B* **164**, 298 (1966).
31. W. H. Li, L. A. Sadler, *Genetics* **129**, 513 (1991).
32. R. Sachidanandam *et al.*, *Nature* **409**, 928 (2001).
33. R. Lewontin, in *Evolutionary Biology* **6**, T. Dobzhansky, M. K. Hecht, W. C. Steere, Eds. (Appleton-Century-Crofts, New York, 1972), pp. 391–398.
34. D. E. Reich, E. S. Lander, *Trends Genet.* **17**, 502 (2001).
35. D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
36. Entrez SNP (www.ncbi.nlm.nih.gov/sites/entrez?db=snp).
37. International HapMap Consortium, *Nature* **437**, 1299 (2005).
38. S. E. Antonarakis, C. D. Boehm, P. J. Giardina, H. H. Kazazian Jr., *Proc. Natl. Acad. Sci. U.S.A.* **79**, 137 (1982).
39. E. S. Lander, D. Botstein, *Cold Spring Harb. Symp. Quant. Biol.* **51**, 49 (1986).
40. B. Kerem *et al.*, *Science* **245**, 1073 (1989).
41. J. Hastbacka *et al.*, *Nat. Genet.* **2**, 204 (1992).
42. L. Kruglyak, *Nat. Genet.* **22**, 139 (1999).
43. K. G. Ardlie, L. Kruglyak, M. Seielstad, *Nat. Rev. Genet.* **3**, 299 (2002).
44. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, *Nat. Genet.* **29**, 229 (2001).
45. N. Patil *et al.*, *Science* **294**, 1719 (2001).
46. S. B. Gabriel *et al.*, *Science* **296**, 2225 (2002); published online 23 May 2002 (10.1126/science.1069424).
47. G. C. Johnson *et al.*, *Nat. Genet.* **29**, 233 (2001).
48. D. E. Reich *et al.*, *Nat. Genet.* **32**, 135 (2002).
49. D. C. Crawford *et al.*, *Nat. Genet.* **36**, 700 (2004).
50. G. A. T. McVean *et al.*, *Science* **304**, 581 (2004).
51. S. A. Tishkoff, B. C. Verrelli, *Annu. Rev. Genomics Hum. Genet.* **4**, 293 (2003).
52. International HapMap Consortium, *Nature* **449**, 851 (2007).
53. J. Sebat *et al.*, *Science* **305**, 525 (2004).
54. A. J. Iafrate *et al.*, *Nat. Genet.* **36**, 949 (2004).
55. E. Tuzun *et al.*, *Nat. Genet.* **37**, 727 (2005).
56. D. A. Hinds, A. P. Kloek, M. Jen, X. Chen, K. A. Frazer, *Nat. Genet.* **38**, 82 (2006).
57. S. A. McCarroll *et al.*, *Nat. Genet.* **38**, 86 (2006).
58. D. P. Locke *et al.*, *Am. J. Hum. Genet.* **79**, 275 (2006).
59. R. Redon *et al.*, *Nature* **444**, 444 (2006).
60. S. A. McCarroll *et al.*, *Nat. Genet.* **40**, 1166 (2008).
61. S. E. Deeb *et al.*, *Nat. Genet.* **20**, 284 (1998).
62. D. Altshuler *et al.*, *Nat. Genet.* **26**, 76 (2000).
63. J. C. Florez, J. N. Hirschhorn, D. Altshuler, *Annu. Rev. Genomics Hum. Genet.* **4**, 257 (2003).
64. I. Pe'er, R. Yelensky, D. Altshuler, M. J. Daly, *Genet. Epidemiol.* **32**, 381 (2008).
65. W. C. Knowler, R. C. Williams, D. J. Pettitt, A. G. Steinberg, *Am. J. Hum. Genet.* **43**, 520 (1988).
66. R. S. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* **52**, 506 (1993).
67. B. Devlin, K. Roeder, *Biometrics* **55**, 997 (1999).
68. J. K. Pritchard, N. A. Rosenberg, *Am. J. Hum. Genet.* **65**, 220 (1999).
69. A. L. Price *et al.*, *Nat. Genet.* **38**, 904 (2006).
70. D. G. Clayton *et al.*, *Nat. Genet.* **37**, 1243 (2005).
71. J. M. Chapman, J. D. Cooper, J. A. Todd, D. G. Clayton, *Hum. Hered.* **56**, 18 (2003).
72. D. A. Hinds *et al.*, *Science* **307**, 1072 (2005).
73. P. I. W. de Bakker *et al.*, *Nat. Genet.* **37**, 1217 (2005).
74. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, *Nat. Genet.* **39**, 906 (2007).
75. K. M. Weiss, J. D. Terwilliger, *Nat. Genet.* **26**, 151 (2000).
76. J. Couzin, *Science* **296**, 1391 (2002).
77. R. J. Klein *et al.*, *Science* **308**, 385 (2005); published online 10 March 2005 (10.1126/science.1109557).
78. A. O. Edwards *et al.*, *Science* **308**, 421 (2005); published online 10 March 2005 (10.1126/science.1110189).
79. J. L. Haines *et al.*, *Science* **308**, 419 (2005); published online 10 March 2005 (10.1126/science.1110359).
80. G. Thorleifsson *et al.*, *Science* **317**, 1397 (2007); published online 9 August 2007 (10.1126/science.1146554).
81. Wellcome Trust Case Control Consortium, *Nature* **447**, 661 (2007).
82. D. Altshuler, M. Daly, *Nat. Genet.* **39**, 813 (2007).
83. E. Zeggini *et al.*, *Nat. Genet.* **40**, 638 (2008).
84. S. Kathiresan *et al.*, *Nat. Genet.* **40**, 189 (2008).
85. C. J. Willer *et al.*, *Nat. Genet.* **40**, 161 (2008).
86. J. C. Barrett *et al.*, *Nat. Genet.* **40**, 955 (2008).
87. G. Lettre *et al.*, *Nat. Genet.* **40**, 584 (2008).
88. M. N. Weedon *et al.*, *Nat. Genet.* **40**, 575 (2008).
89. S. Sanna *et al.*, *Nat. Genet.* **40**, 198 (2008).
90. D. F. Gudbjartsson *et al.*, *Nat. Genet.* **40**, 609 (2008).
91. S. A. McCarroll *et al.*, *Nat. Genet.* **40**, 1107 (2008).
92. C. A. Haiman *et al.*, *Nat. Genet.* **39**, 638 (2007).
93. J. Maller *et al.*, *Nat. Genet.* **38**, 1055 (2006).
94. M. Li *et al.*, *Nat. Genet.* **38**, 1049 (2006).
95. R. R. Graham *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 6758 (2007).
96. R. H. Duerr *et al.*, *Science* **314**, 1461 (2006); published online 26 October 2006 (10.1126/science.1135245).
97. R. Sladek *et al.*, *Nature* **445**, 881 (2007).
98. D. F. Easton *et al.*, *Nature* **447**, 1087 (2007).
99. E. Zeggini *et al.*, *Science* **316**, 1336 (2007); published online 25 April 2007 (10.1126/science.1142364).
100. L. J. Scott *et al.*, *Science* **316**, 1341 (2007); published online 25 April 2007 (10.1126/science.1142382).
101. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes for BioMedical Research, *Science* **316**, 1331 (2007); published online 26 April 2007 (10.1126/science.1142358).
102. A. Rivera *et al.*, *Hum. Mol. Genet.* **14**, 3227 (2005).
103. L. T. Amundadottir *et al.*, *Nat. Genet.* **38**, 652 (2006).
104. M. L. Freedman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14068 (2006).
105. R. McPherson *et al.*, *Science* **316**, 1488 (2007); published online 2 May 2007 (10.1126/science.1142447).
106. N. J. Samani *et al.*, *N. Engl. J. Med.* **357**, 443 (2007).
107. A. Helgadóttir *et al.*, *Science* **316**, 1491 (2007); published online 2 May 2007 (10.1126/science.1142842).
108. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
109. P. R. Burton *et al.*, *Nat. Genet.* **39**, 1329 (2007).
110. M. Cargill *et al.*, *Am. J. Hum. Genet.* **80**, 273 (2007).
111. J. C. Florez *et al.*, *N. Engl. J. Med.* **355**, 241 (2006).
112. N. Grarup *et al.*, *Diabetes* **56**, 3105 (2007).
113. L. Pascoe *et al.*, *Diabetes* **56**, 3101 (2007).
114. R. Saxena *et al.*, *Diabetes* **55**, 2890 (2006).
115. H. Staiger *et al.*, *PLoS One* **2**, e832 (2007).
116. V. Steinthorsdóttir *et al.*, *Nat. Genet.* **39**, 770 (2007).
117. N. Wolf *et al.*, *J. Med. Genet.* **45**, 114 (2008).
118. A. P. Reiner *et al.*, *Am. J. Hum. Genet.* **82**, 1193 (2008).
119. P. M. Ridker *et al.*, *Am. J. Hum. Genet.* **82**, 1185 (2008).
120. W. Winckler *et al.*, *Diabetes* **56**, 685 (2007).
121. J. Gudmundsson *et al.*, *Nat. Genet.* **39**, 977 (2007).
122. G. Thomas *et al.*, *Nat. Genet.* **40**, 310 (2008).
123. S. F. A. Grant *et al.*, *Nat. Genet.* **38**, 320 (2006).
124. A. Hazra *et al.*, *Cancer Causes Control* **19**, 975 (2008).
125. A. R. Folsom *et al.*, *Diabetes Care* **31**, 905 (2008).
126. I. K. Kotowski *et al.*, *Am. J. Hum. Genet.* **78**, 410 (2006).
127. J. Cohen *et al.*, *Nat. Genet.* **37**, 161 (2005).
128. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, *N. Engl. J. Med.* **354**, 1264 (2006).
129. J. N. Hirschhorn, D. Altshuler, *J. Clin. Endocrinol. Metab.* **87**, 4438 (2002).
130. J. C. Cohen *et al.*, *Science* **305**, 869 (2004).
131. W. Ji *et al.*, *Nat. Genet.* **40**, 592 (2008).
132. L. A. Weiss *et al.*, *N. Engl. J. Med.* **358**, 667 (2008).
133. T. Walsh *et al.*, *Science* **320**, 539 (2008); published online 27 March 2008 (10.1126/science.1155174).
134. J. Sebat *et al.*, *Science* **316**, 445 (2007); published online 14 March 2007 (10.1126/science.1138659).
135. J. D. Rioux *et al.*, *Nat. Genet.* **39**, 596 (2007).
136. L. A. Weiss *et al.*, *N. Engl. J. Med.* **358**, 667 (2008).
137. V. Emilsson *et al.*, *Nature* **452**, 423 (2008).
138. 1000 Genomes (www.1000genomes.org).
139. T. J. Albert *et al.*, *Nat. Methods* **4**, 903 (2007).
140. T. S. Mikkelsen *et al.*, *Nature* **448**, 553 (2007).
141. J. Lamb *et al.*, *Science* **313**, 1929 (2006).

Supporting Online Material

www.sciencemag.org/cgi/content/full/322/5903/881/DC1
Table S1

10.1126/science.1156409



Practice of Epidemiology

Gene-Environment Interaction in Genome-Wide Association Studies

Cassandra E. Murcray, Juan Pablo Lewinger, and W. James Gauderman

Initially submitted November 27, 2007; accepted for publication May 5, 2008.

It is a commonly held belief that most complex diseases (e.g., diabetes, asthma, cancer) are affected in part by interactions between genes and environmental factors. However, investigators conducting genome-wide association studies typically test for only the marginal effects of each genetic marker on disease. In this paper, the authors propose an efficient and easily implemented 2-step analysis of genome-wide association study data aimed at identifying genes involved in a gene-environment interaction. The procedure complements screening for marginal genetic effects and thus has the potential to uncover new genetic signals that have not been identified previously.

association; environment; genes; genetic markers; genetics; genome

Abbreviations: E, environment; G, gene; GWAS, genome-wide association study; OR, odds ratio; SNP, single nucleotide polymorphism.

Editor's note: Two invited commentaries on this article appear on pages 227 and 231, and the authors' response is published on page 234.

Many common, complex traits are believed to be a result of the combined effect of genes, environmental factors, and their interactions. For example, Ito et al. (1) showed a significant interaction between smoking status and the apurinic/apyrimidinic endonuclease 1 protein coding gene (*APE1*) for lung cancer. Stern et al. (2) found smoking status to be an effect modifier of the association between the XPD codon 751 polymorphism and risk of bladder cancer. Understanding the relation between genetic polymorphisms and environmental exposures can help to identify high-risk subgroups in the population and provide better insight into pathway mechanisms for complex diseases.

Methods for identifying disease susceptibility genes include linkage analysis, candidate gene association studies, and, more recently, the genome-wide association study (GWAS). It is known that a GWAS can be more powerful than linkage analysis in detecting genes associated with modest increases in disease risk (3). Current GWAS meth-

ods are designed to detect main effects, that is, direct associations of a single nucleotide polymorphism (SNP) or clusters of SNPs with disease (4, 5). In the context of complex diseases, scanning for main effects might miss important genetic variants specific to subgroups of the population, as defined by some exposure. In fact, interactions with opposite effects in 2 different exposure groups (crossing-interaction) will not show a main effect and therefore will not be identified by using standard approaches.

Despite the importance of gene-environment interaction for complex diseases, little work has been done to develop methods for detecting these types of interactions in the context of a GWAS. In this paper, we focus on identifying SNPs that demonstrate heterogeneity between subgroups defined by some environmental exposure. We introduce an efficient 2-step approach for detecting loci involved in gene-environment interactions that is performed independently of any initial scans for main effects. Our method expands on the traditional test for gene-environment interaction in a case-control study by incorporating a preliminary screening step constructed to efficiently use all available information in the data. We demonstrate that this 2-step approach is more powerful than the standard test of interaction across a wide range of models.

Correspondence to Cassandra E. Murcray, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP 222F, Los Angeles, CA 90089-9010 (e-mail: murcray@usc.edu).

MATERIALS AND METHODS

Let D be an indicator of disease status, and assume we have a sample of cases ($D = 1$) and unrelated controls ($D = 0$). Assume that information is available for a binary environmental exposure, with E as an indicator for exposure. Further assume that for each individual we have genotyped M SNPs spanning the genome, with g_1, g_2, \dots, g_M denoting the genotypes at the M loci. Letting G_1, G_2, \dots, G_M denote some genetic coding (e.g., additive, dominant) for each genotype, we consider a model for a given SNP of the form

$$\text{logit } P(D = 1 | g, e) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE. \quad (1)$$

Under a dominant coding of the genotype, for example, $\exp(\beta_g) = \text{OR}_g$ is the odds ratio (OR) comparing carriers of at least 1 risk allele ($G = 1$) with noncarriers ($G = 0$) among those unexposed ($E = 0$). $\exp(\beta_e) = \text{OR}_e$ is the odds ratio comparing risk for exposed ($E = 1$) with that for unexposed ($E = 0$) individuals among noncarriers of the risk allele. Lastly, $\exp(\beta_{ge}) = \text{OR}_{ge}$ is the ratio of the genetic odds ratios comparing exposed with unexposed subjects, that is, $\text{OR}_{g|E=1} / \text{OR}_{g|E=0}$. If this ratio is equal to 1.0, or $\beta_{ge} = 0$, we say that there is no interaction between genotype and the environmental exposure.

In the context of a GWAS, a standard approach to test for $G \times E$ interaction would be to perform a 1-df test of $H_0 : \beta_{ge} = 0$ for each SNP based on the model in equation 1. We assume that a likelihood ratio test will be used to test this hypothesis and denote it as the “1-step” test of interaction. A correction for multiple comparisons (e.g., Bonferroni, controlling the false discovery rate (6)) is required to achieve a desired genome-wide type I error.

One might consider using a case-only analysis of diseased individuals as the sole approach for identifying interactions in a GWAS. Indeed, the case-only test has been shown to be more powerful than a case-control analysis for identifying interactions (7, 8). However, the case-only analysis depends heavily on an underlying assumption of G-E independence in the population, which would be untenable across all SNPs being scanned in a GWAS. If there is an underlying population association between genotype and environmental exposure, a case-only analysis will result in an inflated number of false positives (7, 9).

We propose an alternative 2-step test to scan for interactions that combines the power of the case-only test with the protection from bias of the case-control test. Our 2-step scan for interactions consists of the following 2 steps:

1. Step 1, screening test: For each of the M SNPs, we perform a likelihood ratio test of association between G and E based on the logistic model $\text{logit } P(E = 1 | g) = \gamma_0 + \gamma_g G$. This is the standard test that would be applied in a case-only analysis of $G \times E$ interaction (7, 9), although, in our context, the test is applied to the combined sample of cases and controls. The subset of m SNPs that exceeds a given significance threshold (i.e., with $P < \alpha_1$) for the test of $H_0 : \gamma_g = 0$ is analyzed in step 2.
2. Step 2, case-control test: The m SNPs that pass step 1 are assessed in the traditional test of $G \times E$ interaction, that

is, based on a likelihood ratio test of $H_0 : \beta_{ge} = 0$ derived from the model in equation 1. Significance at this step is defined as having a P value less than α/m , where α is the desired overall type I error rate.

Although step 1 of our procedure is also sensitive to the assumption of independence between G and E in the population, the step 2 comparison of cases to controls is not. Therefore, our overall 2-step procedure will provide a valid test in the presence of population-level association between genotype and exposure, a claim we will verify by simulation.

Given the reported power of a case-only analysis applied to only diseased individuals (7, 9), one may be tempted to apply step 1 to only diseased individuals, that is, to perform a true case-only test of interaction in step 1 and use it to define the subset of m markers to analyze in step 2. However, this approach produces a correlation between the step 1 and step 2 test statistics and leads to an inflated type I error rate for the overall procedure. Our screening test of $G \times E$ interaction applied to the entire sample of cases and controls eliminates the correlation between tests in steps 1 and 2 (Appendix 1) and, as we show (Appendix 2) and verify by simulation, preserves the overall type I error rate.

We performed simulations to study the power achieved by our 2-step testing framework compared with the traditional 1-step method. For each of 1,000 replicate data sets, we simulated a sample of 500 cases and 500 controls, each with genotype information on a large number of markers ($M = 10,000, 25,000, \text{ and } 50,000$). Although larger marker sets are likely to be used in practice, our chosen set sizes are sufficient to demonstrate the relative power of our 2-step method. Markers were assumed to be independent loci distributed across the genome. For each replicate, a single marker was chosen to be the true disease susceptibility locus, with remaining markers assumed to have no association with disease. We considered a range of minor allele frequencies (q_A) for the disease susceptibility locus, including 0.10, 0.20, and 0.30. For the remaining null markers, we simulated a uniform distribution of allele frequencies between 0.05 and 0.30. We also considered a range of values for the exposure prevalence (p_E), including 0.10, 0.25, and 0.50, and set the population disease prevalence to 0.05. Finally, we considered a range of possible values for the genetic and environmental main effects (β_g and β_e , respectively) as well as for the interaction effect (β_{ge}).

As described above, the traditional case-only method of testing for $G \times E$ interaction is based on the assumption of no population-level association between the G and E . We explored the sensitivity of our method to population-level association between G and E by introducing a parameter p_{ge} , defined to be the probability that a given marker is associated with the exposure at the population level. With $p_{ge} = 0.0$, none of the markers was assumed to be associated with E in the population. It is unlikely that a large percentage of SNPs will be associated with a given environmental factor, but, for completeness, we considered a wide range of values—0.01 to 0.95. For each simulated marker, we randomly decided whether it was associated with E in the population based on probability p_{ge} . For any marker

chosen to have an association, we generated genotypes conditional on the assigned E and an assumed population marker–exposure odds ratio of 2.0.

For each parameter setting, we applied the traditional 1-step G × E test and our 2-step approach. We estimated the experiment-wise type I error as the proportion of 1,000 replicates in which at least one of the null markers was found to be significant after a Bonferroni correction for multiple comparisons. Power was calculated as the number of replicates in which the disease susceptibility locus was detected at an overall significance level of 0.05, again after a Bonferroni correction for multiple comparisons. We also computed the proportion of replicates in which the disease susceptibility locus was among the top 10 or top 25 most significant SNPs, that is, on a short list that might warrant additional scrutiny following a genome-wide scan.

RESULTS

Across a range of interaction effect sizes ($R_{ge} = \exp(\beta_{ge})$), our 2-step method was more powerful than the standard 1-step test for detecting an interaction (Figure 1). For example, when $R_{ge} = 3.0$, power was 33.2% using a standard 1-step approach compared with 57.9% using our 2-step method. As we would expect, as the effect size of the interaction increases, both tests gain power. For a small interaction effect, both tests have low power to detect a causal locus; at a sufficiently large effect size, the 2 tests approach 100% power. The largest differences in power between the 2 methods occurred when the interaction effect was of moderate magnitude, from 2.5 to 4.0. The estimates of power in Figure 1 all assumed a disease susceptibility locus allele frequency $q_A = 0.2$, exposure frequency $p_E = 0.5$, no main effects ($R_g = R_e = 1$), no population-level association between g and E ($p_{ge} = 0$), 10,000 SNP markers, and a first-step significance threshold of $\alpha_1 = 0.05$.

For various alternatives to the above parameter settings, Table 1 shows type I error and power for the 1- and 2-step methods for detecting G-E interaction, holding the interaction effect size fixed ($R_{ge} = 3.0$). Both methods approximated the nominal 0.05 type I error under all scenarios, even when there was a population-level association ($p_{ge} \neq 0$) between markers and the environmental factor (also refer to the Web Figure, which is posted on the *Journal's* website (<http://aje.oupjournals.org/>)). The mean type I error across all scenarios was slightly smaller for the 2-step test (mean = 0.051) than for the conventional 1-step test (mean = 0.056), indicating that, on average, the 2-step test was slightly more conservative than the 1-step test. The 2-step test was consistently more powerful than the traditional 1-step case-control test over a wide variety of parameter settings. As expected, power for both tests was highest for common exposures and alleles. The 2-step method was at least twice as powerful as the 1-step test when the exposure was rare or when the disease allele was rare, although absolute power in these situations was low for both procedures. Power for the 2-step test depended somewhat on the significance threshold for step 1 (α_1). Specifically, relative to our base model with $\alpha_1 = 0.05$, a smaller threshold value ($\alpha_1 = 0.01$) resulted in increased power to detect the disease susceptibility locus,

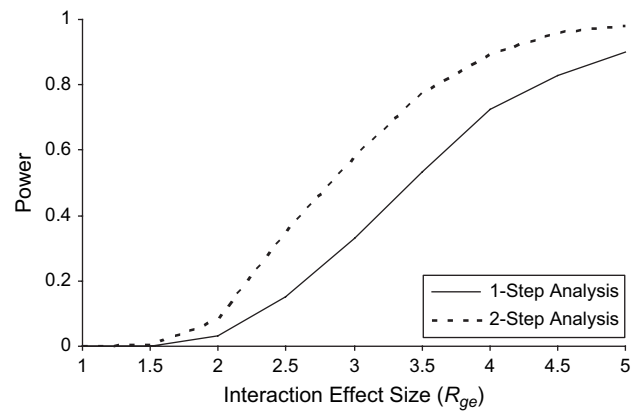


Figure 1. Power for 1-step and 2-step analyses for varying levels of interaction effect size (R_{ge}). All other parameter settings remain constant under “base” model specifications ($M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $p_{ge} = 0$, $\alpha_1 = 0.05$).

while we saw reduced power when we allowed more markers to move into step 2 ($\alpha_1 = 0.10$).

As expected, a population-level association between markers and environment ($p_{ge} > 0$) increased the number of markers that proceeded to step 2. However, for the range of values we considered plausible in a genome-wide scan ($p_{ge} = 0.01$ or 0.05), there was not an appreciable impact on power using the 2-step method. At more liberal values for the proportion of markers with a population-level association between G and E ($p_{ge} = 0.30, 0.95$), power for the 2-step method approached that for the traditional 1-step method. Specifically, when we assumed that 95% of the markers were associated with the environmental factor, power estimates for the 1- and 2-step methods were identical (29.3%). Power estimates under the full range of values for p_{ge} we considered are shown in the Web Figure.

Across a range of interaction effect sizes ($R_{ge} = \exp(\beta_{ge})$), our 2-step method was also more likely than the 1-step method to detect the disease susceptibility locus according to the ranked P -value statistics (Figure 2). Using the ordered P values compared with a traditional significance threshold to choose a subset of highly ranked SNPs did not depend as much on interaction effect size to be confident that the selected subset included the disease susceptibility locus. For example, our 2-step approach resulted in the disease susceptibility locus being in the top-10-ranked P values in 79% of the 1,000 replicates with an interaction effect size of 2.5 compared with 59% when the 1-step method was used. Power for this interaction effect size was less than 35% for both tests (Figure 1).

Our 2-step method was more robust to changes in exposure prevalence, minor allele frequency, number of markers, and so forth, when we focused on the rank statistics (Table 2). Under our base model settings, the disease susceptibility locus was in the top 10 P values in 94% of the 1,000 replicates using the 2-step method compared with only 79% of the replicates when the 1-step method was used. Even under the least ideal circumstances, with a low exposure

Table 1. Type I Error and Power for 1-Step and 2-Step Tests for Gene-Environment Interaction

Model ^a	Type I Error ^b		Power ^c	
	1 Step	2 Step	1 Step	2 Step
Base ^d	0.062	0.045	0.332	0.579
Disease susceptibility locus allele frequency (q_A)				
0.1	0.062	0.043	0.158	0.348
0.3	0.052	0.039	0.344	0.588
Exposure prevalence (p_E)				
0.1	0.058	0.073	0.035	0.111
0.25	0.053	0.052	0.212	0.448
Effect sizes (R_g, R_e, R_{ge})				
123	0.063	0.038	0.228	0.479
213	0.054	0.054	0.312	0.578
223	0.053	0.054	0.253	0.483
Population gene-environment association (p_{ge})				
0.01	0.054	0.040	0.320	0.564
0.05	0.057	0.050	0.316	0.504
0.30	0.061	0.049	0.315	0.385
0.95	0.056	0.052	0.293	0.293
No. of markers (M)				
25,000	0.051	0.063	0.290	0.518
50,000	0.049	0.060	0.223	0.436
Step 1 threshold (α_1)				
0.01	0.065	0.049	0.350	0.698
0.1	0.051	0.048	0.330	0.500

^a Variations to model specification refer to a change to a parameter setting in the “base” model. All other parameters remain constant.

^b All estimates of type I error have a standard error of <0.008.

^c All estimates of power have a standard error of <0.016.

^d $M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_{ge} = 0$, $\alpha_1 = 0.05$.

prevalence, the ranked P value for the disease susceptibility locus was still in the top 25 in 63% of replicates using the 2-step method compared with 35% when the 1-step method was used. Under all other scenarios we examined, the ranked P value using the 2-step method was in the top 25 for at least 85% of replicates, and the 2-step method always outperformed the 1-step approach.

DISCUSSION

For a GWAS in a case-control sample, we have shown that our 2-step testing approach provides a powerful alternative for testing $G \times E$ interaction relative to a traditional 1-step test. For the parameter settings we examined, the 2-step method was always more powerful than the 1-step method. Our method was also more robust than the traditional case-control test to changes in allele frequency, exposure prevalence, and other parameters when comparing the ranked

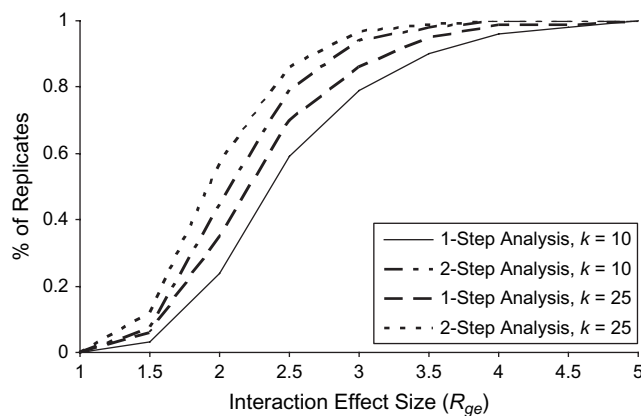


Figure 2. Percentage of replicates for which the P value for disease susceptibility locus is ranked in the top k ($k = 10$ or $k = 25$) marker P values for varying levels of interaction effect size (R_{ge}). All other parameter settings remain constant under “base” model specifications ($M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $p_{ge} = 0$, $\alpha_1 = 0.05$).

P value for the true disease-susceptibility locus. Given its increased power and ease of implementation, the 2-step test is an attractive alternative for identifying $G \times E$ interactions in a GWAS for complex diseases.

We assumed a dichotomous environmental factor and dominant susceptibility locus in our simulations. In practice, the investigator may be interested in an alternative type of environmental factor (e.g., continuous, multinomial) or genetic model (e.g., additive, codominant). Both steps in our method can naturally be extended to accommodate any parameterization of the environmental exposure or genetic coding. The absolute power of these extensions would depend on the underlying data distributions, but we would expect similar increases in power for our 2-step approach relative to a 1-step approach.

A typical GWAS is conducted on a large sample size to achieve power to detect modest-sized effects at genome-wide significance after correction for multiple testing. We considered a scenario with only 500 cases and controls genotyped on 10,000 markers for our base model. With an increase in sample size, power to detect a marker involved in a $G \times E$ interaction would increase for both the 1- and 2-step methods. An increase in number of markers could increase or decrease power, depending on whether the increase in linkage disequilibrium between the disease susceptibility locus and the markers offsets the penalty for a larger number of tests. However, we would expect variations in sample size and number of markers to affect both the 1- and 2-step approaches similarly and therefore not affect the relative comparison of power for the 2 methods.

We simulated scenarios in which a proportion (p_{ge}) of the available null markers were associated with the environmental factor in the population to establish that our 2-step approach preserves the desired type I error rate. It is possible for a causal locus to influence disease risk through an interaction with some environmental factor and simultaneously to be associated with the same environmental exposure in the

Table 2. Percentage of Replicates With a Ranked *P* Value for Disease Susceptibility Locus in the Top 10 or Top 25 Single Nucleotide Polymorphisms for 1-Step and 2-Step Tests for Gene-Environment Interaction

Model ^a	Top 10		Top 25	
	1 Step	2 Step	1 Step	2 Step
Base ^b	0.79	0.94	0.86	0.97
Disease susceptibility locus allele frequency (q_A)				
0.1	0.57	0.85	0.68	0.91
0.3	0.78	0.93	0.86	0.95
Exposure prevalence (p_E)				
0.1	0.26	0.51	0.35	0.63
0.25	0.66	0.89	0.76	0.94
Effect sizes (R_G, R_E, R_{GE})				
123	0.71	0.91	0.81	0.95
213	0.76	0.93	0.84	0.96
223	0.68	0.91	0.79	0.94
Population gene-environment association (p_{ge})				
0.01	0.76	0.94	0.85	0.97
0.05	0.79	0.91	0.86	0.94
0.30	0.75	0.84	0.84	0.91
0.95	0.75	0.76	0.84	0.85
No. of markers (M)				
25,000	0.69	0.93	0.78	0.95
50,000	0.65	0.87	0.74	0.92
Step 1 threshold (α_1)				
0.01	0.81	0.92	0.88	0.93
0.1	0.76	0.91	0.85	0.95

^a Variations to model specification refer to a change to a parameter setting in the “base” model. All other parameters remain constant.

^b $M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_E = 0.5$, $R_G = 1$, $R_E = 1$, $R_{GE} = 3$, $p_{ge} = 0$, $\alpha_1 = 0.05$.

general population. Under this scenario, the population-level association between the causal locus and the environmental factor would affect the power of our screening step. For example, a population-level G-E association at the causal locus in the opposite direction of the G × E interaction effect may reduce the power of our method. On the other hand, a positive G × E interaction combined with a positive G-E association in the population would inflate the estimated G-E association in our screening step and would likely increase the overall power of our 2-step test.

Incorporating a screening step to improve power in genetic analysis has been proposed in other contexts. For example, Van Steen et al. (10) developed a 2-step method for genome-wide association tests of genetic main effects using family data. Their screening step was based on a regression model using between-family information, and it was statistically independent of the family-based association test used in the second analysis step. Similar to their approach, our proposed analysis begins with a potentially biased test in the

first step that is designed to efficiently screen for potentially important SNPs and then uses an unbiased second step to ensure an overall valid procedure. Millstein et al. (11) also used a screening step in their Focused Interaction Testing Framework software to identify genes involved in G × G interactions in a study of many candidate loci. Although this screening test was biased in the presence of population-level association among genes, their second-step model ensured that the overall Focused Interaction Testing Framework approach was unbiased. Our results, in combination with those of Van Steen et al. and Millstein et al., demonstrate that well-designed 2-step approaches can lead to improved power in a wide range of genetic applications.

The additional power of our 2-step procedure comes from exploiting independent information provided by oversampling of cases relative to their prevalence in the population. In the presence of G × E interaction, this oversampling of cases induces an association between G and E in the combined case-control sample. Although it would be possible to develop an alternative 1-step test based on a likelihood that incorporates this additional information, such a test would not preserve the type I error in the presence of population-level G-E association. In the GWAS context, however, we can use the additional information derived from the oversampling of cases in a screening step to reduce the number of SNPs to be tested in the second step. When the power of the first step test is high, the chance that a true positive will be carried to the second step is also high. At the same time, a large number of null SNPs will be eliminated by the first-step screen, which reduces the multiple testing burden and results in our observed, overall gain in power to detect interaction at the causal locus.

Our 2-step approach is preferable to a case-only analysis of affected individuals, since the latter will have an inflated type I error rate in the presence of population G-E association. Even if a small subset of null SNPs has a population G-E association, one would expect several thousand false-positive results using a case-only analysis of interaction given the overall number of SNPs being tested. On the other hand, the type I error of our 2-step procedure is maintained because the second-step test is unbiased and the 2 steps are independent. Thus, even if there is a strong association in the population between E and a specific null SNP such that the SNP passes our first-step screen, the type I error rate for our overall 2-step procedure will be maintained.

Kraft et al. (12) proposed a powerful 2-df test for assessing genetic main effects and interactions jointly. They showed that under a wide variety of parameter settings, the 2-df test was often more powerful than a test of the main effect or the traditional test for G × E interaction. Although it is possible that a 2-df test would be more powerful than our 2-step testing framework, many investigators conducting a GWAS will want to begin with a full scan for genetic main effects. After conducting such a scan, use of the 2-df test would assess redundant information through partial retesting of the main effect. If a secondary goal is to detect genes involved in G × E interactions, our method allows it to be performed independently of the main-effect scan.

Several reported genome-wide association studies have conducted an initial scan of all available genotypes for main

effects by ignoring heterogeneity between exposure classifications (13–17). It is possible that by focusing completely on main effects, SNPs with disease associations modified by some environmental factor were not detected. Still, it has been argued that focusing on $G \times E$ interaction is not advantageous over testing for main effects (18). However, if there is strong evidence that an environmental factor contributes to risk and possibly modifies a genetic effect, it is potentially important to define a testing strategy that uses independent information in a second scan across the available markers. We developed this method in order to use this additional information to detect genetic heterogeneity across subgroups that might otherwise be missed in a direct main effect test. Although we focused on subgroups defined by an environmental factor, our approach can be used to assess heterogeneity across racial/ethnic groups, genotypes at another locus ($G \times G$ interaction), or any other variable that can be used to classify study subjects.

Our method relies on a priori knowledge of factors that might be expected to modify the risk of genotype on disease. For example, the Children's Health Study, a prospective cohort study designed to investigate respiratory outcomes in children in 12 communities throughout southern California, has shown evidence to suggest that both regional air quality and proximity to traffic contribute to risk of asthma, reduced lung function growth, and other respiratory outcomes (19–23). For the GWAS being conducted in this cohort, simply a scan of the main effects that ignored the potential modification of genetic effects by air pollutants might lead investigators to miss genetic variants that are important determinants of complex respiratory diseases.

We have shown in the context of a GWAS that utilizing ascertainment information through a screening step of available markers can lead to substantial increases in power to detect a gene involved in a $G \times E$ interaction. We furthermore showed that this 2-step approach is more robust to changes in environmental exposure, minor allele frequency, and so forth, than the traditional 1-step test for identifying highly ranked SNPs. Our approach therefore has the potential to increase the yield of a given GWAS by identifying additional, important loci that act in concert with an environmental or other factor to influence risk of a complex disease.

ACKNOWLEDGMENTS

Authors affiliation: Department of Preventive Medicine, University of Southern California, Los Angeles, California (Cassandra E. Murcray, Juan Pablo Lewinger, W. James Gauderman).

This work was supported in part by the National Institute of Environmental Health Sciences (T32 ES013678, 5P30ES007048, 5P01ES009581, R826708, RD831861, 5P01ES011627, 5R01ES014447, and 5R01ES014708), the National Heart, Lung, and Blood Institute (5R01HL087680, 5R01HL61768, and 5R01HL76647), the National Human Genome Research Institute (P50 HG 002790-02), and the Hastings Foundation.

Conflict of interest: none declared.

REFERENCES

- Ito H, Matsuo K, Hamajima N, et al. Gene-environment interactions between the smoking habit and polymorphisms in the DNA repair genes, APE1 Asp148Glu and XRCC1 Arg399Gln, in Japanese lung cancer risk. *Carcinogenesis*. 2004;25(8):1395–1401.
- Stern MC, Johnson LR, Bell DA, et al. XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2002;11(10 pt 1):1004–1011.
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516–1517.
- Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol*. 2007;31(5):365–375.
- Zhao J, Jin L, Xiong M. Nonlinear tests for genomewide association studies. *Genetics*. 2006;174(3):1529–1538.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc (B)*. 1995;57:289–300.
- Khouri MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol*. 1996;144(3):207–213.
- Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol*. 2002;155(5):478–484.
- Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153–162.
- Van Steen K, McQueen MB, Herbert A, et al. Genomic screening and replication using the same data set in family-based association testing. *Nat Genet*. 2005;37(7):683–691.
- Millstein J, Conti DV, Gilliland FD, et al. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet*. 2006;78(1):15–27.
- Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007;63(2):111–119.
- Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007;447(7148):1087–1093.
- Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*. 2007;39(7):870–874.
- Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007;316(5829):1331–1336.
- Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*. 2007;316(5829):1341–1345.
- Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007;316(5829):1336–1341.
- Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*. 2001;358(9290):1356–1360.

19. Gauderman WJ, Avol E, Gilliland F, et al. The effect of air pollution on lung development from 10 to 18 years of age. *N Engl J Med.* 2004;351(11):1057–1067.
 20. Gauderman WJ, Avol E, Lurmann F, et al. Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology.* 2005;16(6):737–743.
 21. Gauderman WJ, Vora H, McConnell R, et al. Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. *Lancet.* 2007;369(9561):571–577.
 22. McConnell R, Berhane K, Gilliland F, et al. Air pollution and bronchitic symptoms in Southern California children with asthma. *Environ Health Perspect.* 1999;107(9):757–760.
 23. McConnell R, Berhane K, Gilliland F, et al. Asthma in exercising children exposed to ozone: a cohort study. *Lancet.* 2002;359(9304):386–391.
 24. van der Vaart AW. *Asymptotic Statistics.* Cambridge, United Kingdom: Cambridge University Press; 1998.

APPENDIX 1

The Step-1 and Step-2 Test Statistics Are Asymptotically Uncorrelated

We assume a binary exposure E and binary genotype G , although the following result can be readily extended to multilevel exposures and genotypes. Let us consider the $2 \times 2 \times 2$ table for a case-control study in which, as before, D is the binary case-control indicator:

	$D = 1$	
	$G = 1$	$G = 0$
$E = 1$	n_{11}	n_{12}
$E = 0$	n_{13}	n_{14}

	$D = 0$	
	$G = 1$	$G = 0$
$E = 1$	n_{01}	n_{02}
$E = 0$	n_{03}	n_{04}

The standard interaction $G \times E$ odds ratio is $OR_2 = (n_{11}n_{14}/n_{12}n_{13})/(n_{01}n_{04}/n_{02}n_{03})$, and the $G \times E$ odds ratio pooling cases and controls is $OR_1 = (n_{11} + n_{01})(n_{14} + n_{04})/(n_{12} + n_{02})(n_{13} + n_{03})$. $\log(OR_1)$ is the numerator of a Wald test statistic for step 1, and $\log(OR_2)$ is the numerator of a Wald statistic for step 2. We show that the asymptotic covariance $\text{Cov}(\log(OR_1), \log(OR_2)) = 0$ by using the delta method. The delta method establishes that if a random vector \mathbf{T}_n is asymptotically multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $n \rightarrow \infty$, then a differentiable transformation $f(\mathbf{T}_n)$ is asymptotically multivariate normal $N(f(\boldsymbol{\mu}), Df(\boldsymbol{\mu})^T \boldsymbol{\Sigma} Df(\boldsymbol{\mu}))$, where Df is the matrix of first-order derivatives of f (24).

The vectors $\mathbf{Y} = (n_{11}, n_{12}, n_{13}, n_{14})$ and $\mathbf{Z} = (n_{01}, n_{02}, n_{03}, n_{04})$ of observed counts in cases and controls are independent and have a multinomial distribution $\text{Mult}(p_1, p_2, p_3, p_4, n_1)$ and $\text{Mult}(q_1, q_2, q_3, q_4, n_0)$, respectively, where n_1 is the number of cases; n_0 is the number of controls; p_1, p_2, p_3, p_4 are the cell frequencies of the 2×2 $G \times E$ table for the cases; and q_1, q_2, q_3, q_4 are the cell frequencies of the 2×2 $G \times E$ table for the controls. By the standard approximation to the multinomial distribution, $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ is asymptotically normal with mean $\boldsymbol{\mu} = E[\mathbf{X}] = (n_1 p_1, n_1 p_2, n_1 p_3, n_1 p_4, n_0 q_1, n_0 q_2, n_0 q_3, n_0 q_4)$ and partitioned covariance matrix

$$\Sigma = \left[\begin{array}{cccc|cccc} n_1 p_1 (1 - p_1) & -n_1 p_1 p_2 & -n_1 p_1 p_3 & -n_1 p_1 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_2 & n_1 p_2 (1 - p_2) & -n_1 p_2 p_3 & -n_1 p_2 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_3 & -n_1 p_2 p_3 & n_1 p_3 (1 - p_3) & -n_1 p_3 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_4 & -n_1 p_2 p_4 & -n_1 p_2 p_3 & n_1 p_4 (1 - p_4) & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & n_0 q_1 (1 - q_1) & -n_0 q_1 q_2 & -n_0 q_1 q_3 & -n_0 q_1 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_2 & n_0 q_2 (1 - q_2) & -n_0 q_2 q_3 & -n_0 q_2 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_3 & -n_0 q_2 q_3 & n_0 q_3 (1 - q_3) & -n_0 q_3 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_4 & -n_0 q_2 q_4 & -n_0 q_2 q_3 & n_0 q_4 (1 - q_4) \end{array} \right]$$

If $f(\mathbf{X}) = (\log(OR_2(\mathbf{X})), \log(OR_1(\mathbf{X})))$, we have

$$Df(\boldsymbol{\mu}^T) = \left[\begin{array}{cccccccc} \frac{1}{n_1 p_1} & \frac{-1}{n_1 p_2} & \frac{-1}{n_1 p_3} & \frac{1}{n_1 p_4} & \frac{-1}{n_0 q_1} & \frac{1}{n_0 q_2} & \frac{1}{n_0 q_3} & \frac{-1}{n_0 q_4} \\ \frac{1}{n_1 p_1 + n_0 q_1} & \frac{-1}{n_0 p_2 + n_0 q_2} & \frac{-1}{n_1 p_3 + n_0 q_3} & \frac{1}{n_1 p_4 + n_0 q_4} & \frac{1}{n_1 p_1 + n_0 q_1} & \frac{-1}{n_0 p_2 + n_0 q_2} & \frac{-1}{n_1 p_3 + n_0 q_3} & \frac{1}{n_1 p_4 + n_0 q_4} \end{array} \right]$$

and by the delta method, the asymptotic covariance matrix of $f(\mathbf{X})$ is

$$Df(\boldsymbol{\mu})^T \Sigma Df(\boldsymbol{\mu}) = \begin{bmatrix} \frac{1}{n_1 p_1} + \frac{1}{n_1 p_2} + \frac{1}{n_1 p_3} + \frac{1}{n_1 p_4} + \frac{1}{n_0 q_1} + \frac{1}{n_0 q_2} + \frac{1}{n_0 q_3} + \frac{1}{n_0 q_4} & 0 \\ 0 & g(p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) \end{bmatrix},$$

where $g = (n_1, n_0, p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4)$ is a rather lengthy expression involving $n_1, n_0, p_1, p_2, p_3, p_4, q_1, q_2, q_3,$ and q_4 . The asymptotic joint distribution of $\log(\text{OR}_1)$ and $\log(\text{OR}_2)$ is then bivariate normal with covariance given by the off-diagonal entry of the 2×2 matrix above, that is, zero. Note that this result holds for *any* values of the cell frequencies and thus for any values of the underlying model parameters. This establishes the asymptotic independence of the 2 Wald statistics and, because they are asymptotically equivalent, of the corresponding likelihood ratio test statistics that we propose.

APPENDIX 2

The 2-Step Test Preserves the Type I Error

Let M be the total number of SNP markers, \mathcal{M}_0 be the subset of SNPs that are true negatives, \mathcal{M}_1 be the subset of SNPs that are true positives (i.e., for which there is $G \times E$ interaction), and M_0 and M_1 be the number of SNPs in \mathcal{M}_0 and \mathcal{M}_1 , respectively. Let T_{1k}, T_{2k} be the first- and second-step likelihood ratio statistics for SNP k , $1 \leq k \leq M$, α the desired genome-wide type I error level, and $\alpha_1 \approx \Pr(T_{1k} > c_1 | H_0)$ the step-1 type I error.

By a standard Bonferroni inequality argument, the genome-wide type I error is guaranteed to be less than α if the critical values c_1 and c_2 are chosen so that $\Pr(T_{1k} > c_1, T_{2k} > c_2) \leq \alpha/M_0$ for $k \in \mathcal{M}_0$. Because T_{1k} and T_{2k} are (asymptotically) independent (Appendix 1), it is equivalent to requiring

$$\Pr(T_{1k} > c_1) \leq \frac{\alpha}{M_0 \times \Pr(T_{1k} > c_1)} \approx \frac{\alpha}{M_0 \alpha_1}. \quad (\text{A1})$$

Now, the number of true negative SNPs tested in the second step can be written as $m_0 = \sum_{k \in \mathcal{M}_0} I(T_{1k} > c_1) = \sum_{k \in \mathcal{M}_0} I_k$, where $I_k = I(T_{1k} > c_1)$ and $I(\cdot)$ is the indicator function. This is a sum of M_0 nonindependent (because of linkage disequilibrium) Bernoulli random variables with probability of success $\approx \alpha_1$. The expected number of true negative SNPs to be tested in step 2 is therefore $E[m_0] \approx M_0 \alpha_1$, which is the denominator on the right-hand side of equation A1. For small α_1 , $\text{Var}(m_0) = \text{Var}(\sum_{k \in \mathcal{M}_0} I_k)$ is small (relative to M_0) because $\text{Var}(I_k) \approx \alpha_1(1 - \alpha_1)$, $\text{Cov}(I_i, I_j) \approx 0$ for SNP pairs in low linkage disequilibrium with each other (the vast majority), and $|\text{Cov}(I_i, I_j)| \leq \alpha_1(1 - \alpha_1)$ for pairs of SNPs in linkage disequilibrium with each other. Therefore, m_0 is close to its expectation $E[m_0]$ with high probability, and the right-hand side of equation A1, $\alpha/M_0 \alpha_1 \approx \alpha/E[m_0]$, is in turn approximately α/m_0 . Since the total number of SNPs tested in the second step is $m \geq m_0$, it suffices to choose c_2 as the $1 - \alpha/m$ quantile of a chi-square distribution with 1 df to satisfy equation A1 or, equivalently, to reject the null hypothesis if the second-step P value is smaller than α/m .



Invited Commentary

Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies—Challenges and Opportunities

Muin J. Khoury and Sholom Wacholder

Initially submitted May 30, 2008; accepted for publication July 25, 2008.

The recent success of genome-wide association studies in finding susceptibility genes for many common diseases presents tremendous opportunities for epidemiologic studies of environmental risk factors. Analysis of gene-environment interactions, included in only a small fraction of epidemiologic studies until now, will begin to accelerate as investigators integrate analyses of genome-wide variation and environmental factors. Nevertheless, considerable methodological challenges are involved in the design and analysis of gene-environment interaction studies. The authors review these issues in the context of evolving methods for assessing interactions and discuss how the current agnostic approach to interrogating the human genome for genetic risk factors could be extended into a similar approach to gene-environment-wide interaction studies of disease occurrence in human populations.

environment; epidemiologic methods; genetics; genomics

Abbreviations: GEWIS, gene-environment-wide interaction studies; GWAS, genome-wide association studies; HuGE, human genome epidemiology.

The breakthrough of this year has to do with humans, genomes, and genetics. But it is not about THE human genome (as if there were only one!). Instead, it is about your particular genome, or mine, and what it can tell us about our backgrounds and the quality of our futures. (1)

For human genomics research, 2007 was a banner year. Using genome-wide analytic platforms that can measure hundreds of thousands of genetic variants simultaneously, more than 100 epidemiologic studies have uncovered genetic risk factors for a wide variety of common, complex diseases (2). Human geneticists are anxious to reap the benefits of the human genome project (3) and the international HapMap project (4) by integrating genomics into health care and disease prevention. Genome-wide association studies (GWAS) have shown that an “agnostic” approach can interrogate the totality of the human genome and identify genetic variants associated with numerous diseases.

Certainly, the functional and clinical implications of the loci detected by using GWAS are far from clear, just as some

of the rare, high-penetrance genetic variants for breast cancer, such as *BRCA1* and *BRCA2*. Biologic studies that assess the role of these variants in disease processes and risk factors may give epidemiologists clues about environmental exposures likely to be involved in human diseases (5, 6). So far, however, the odds ratios of individual genetic variants detected are small, mostly between 0.67 and 1.5 (2), and may not be useful for clinical prediction (7, 8).

A NEW ERA OF GENE-ENVIRONMENT-WIDE INTERACTION STUDIES (GEWIS)

The availability of GWAS and their rapidly declining prices, together with the emergence of collaborative epidemiologic consortia and networks (9, 10), offer major opportunities to epidemiologic researchers focused on effects of the environment, broadly defined to include behavioral, chemical, physical, and social factors (11, 12). The increasing rate of published studies focusing on gene-environment interactions pales against the exploding acceleration in published reports of

Table 1. Trends in Published HuGE Articles, GWAS, and Studies Reporting on GEI, 2001–2007^a

Year	Total HuGE Articles, no.	GWAS		GEI	
		No.	%	No.	%
2001	2,488	0	0	373	15.4
2002	3,196	0	0	444	13.9
2003	3,474	3	0.1	447	12.9
2004	4,279	0	0	518	12.1
2005	5,028	5	0.1	706	14.0
2006	5,357	12	0.2	727	13.6
2007	7,168	105	1.5	1,016	14.2

Abbreviations: GEI, gene-environment interactions; GWAS, genome-wide association studies; HuGE, human genome epidemiology.

^a Data were derived from the HuGE Navigator (13), searched on-line July 10, 2008 at <http://www.hugenavigator.net/>.

genetic association studies. Table 1 shows the trends in published genetic association articles from 2001 to 2007, as captured by the HuGE Navigator (13), an online curated and searchable knowledge base in human genome epidemiology (HuGE), sponsored by the Human Genome Epidemiology Network (HuGENet (14)). Between 2001 and 2007, the number of total HuGE articles almost tripled and the number of reported GWAS articles rose from 0 to more than 150; the number of articles reporting on gene-environment interaction also increased, but the proportion of such articles in the total HuGE literature remained relatively flat at about 14%. We do note that undoubtedly a substantial fraction of nonsignificant tests of gene-environment interaction are unreported, leading to a distortion of the literature and too much attention to the positive reports (15). This possibility, however, is unlikely to affect the literature trends unless publication bias varies with time.

The paucity of established gene-environment interactions to date (refer to García-Closas et al. (16) for a rare exception) in the face of substantial investment in the effort should not overly discourage epidemiologists. The “candidate-gene” approach to studies of genetic factors failed to find and replicate many associations, probably because genetic epidemiologists overestimated their ability to select the best candidates and because the threshold for claiming an association was too low given the low prior probability for even the best candidates (17). But just as the GWAS approach, with its broad interrogation of the genome and rigorous threshold for calling effects significant, identified common variants associated with common disease, so too are GWAS likely to help identify genetic factors that interact with environmental factors.

We have known for decades that failure to incorporate both genetic and environmental factors in a joint analysis will weaken the observed associations between a true risk factor and disease occurrence. Because the pools of susceptible and nonsusceptible persons are mixed, the observed associations tend to be shifted toward the null (18). Theoretically, if we are able to measure gene-environment interactions, we should sharpen our measurements of effects in subsets of the population and even potentially increase our statistical power in measuring such effects (19).

OBSTACLES IN ASSESSING GENE-ENVIRONMENT INTERACTION

There are obstacles on the “environment” side of gene-environment interaction that are not present on the “gene” side. Environmental epidemiology does not have the economy of scale seen in genomics, where the difference in cost between measuring a million variants and one variant is a small fraction of the average cost per participant in a case-control or cohort study that collects DNA. We may be missing important environmental determinants of disease because we do not know what to look for or because we do not know how or when to measure accurately what we do know to seek. A person’s genetic makeup may be too far removed from complex physiologic or biochemical processes that could be more important risk factors for disease. Germ-line variation is static and so can be captured at any point, but variation in the timing of exposure, and the timing of subsequent risk, complicates study of environmental factors; at the same time, variation in exposure and risk over time can provide important clues about etiology. In addition, the major advances in the use of biomarkers in research and medical applications, most notably for infectious diseases, are not yet close to yielding useful measures of long-term exposure regarding diet, pharmaceuticals, and polluted air and water for the large numbers of persons needed for studies of rare diseases. Even as biomarkers continue to improve measurement of some exposures, we must also improve the accuracy of epidemiologic questionnaires, medical records, occupational records, and other proxy measurements of environmental factors.

Investigation of gene-environment interaction to learn about etiology and public health is feasible with existing data. An agnostic strategy that is implemented carelessly, however, will generate a large supply of false-positive findings and cause well-founded skepticism about claims of interactions, given the low prior probabilities of most hypotheses (15). Researchers conducting GWAS are demanding replications and requiring *P* values for significance below what we have ever thought realistic in epidemiology (20) in order to avoid false-positive findings in studying main effects of a million genetic variants. Imagine 10–30 times more tests of interaction involving genes, demographic factors, and personal and environmental exposures. Hypotheses about interaction have lower prior probabilities and tests have lower power for detecting interactions compared with tests for main effects with comparable effect size. In addition, exposures are measured with more significant misclassification than genetic variants are. Huge sample sizes are required to reach the very low *P* values for GWAS of main effect that are finding small effects. How will we decide on and achieve the enormous sample sizes needed for interactions when there are more hypotheses and lower prior probabilities of effect, and when good exposure assessment will be critical? How will we be able to distinguish and draw attention to the few interactions likely to be real from the myriad of false-positive ones?

The decades-old problem of defining interaction (21, 22) is even more prominent in the GWAS era. The statistical models we have used to declare interaction as departure from additive or multiplicative joint effects may be inadequate

to describe the underlying biology of joint gene-environment effects on complex disease. The flood of new empirical data becoming available may allow us to examine both gene-gene and gene-environment interactions in new ways.

Systems biology provides novel experimental approaches to quantify molecular components of a biologic system, to assess their interactions, and to integrate such information into graphic models that may explain or predict emergent phenomena (23). However, there is still a large schism between modeling of interactions in cellular and biologic processes and our ability to use that information in observing health and disease in human populations. How can we use biologic information for defining interaction or choosing which analytic method is most useful for identifying risk factors, genetic or environmental; for describing their joint effects; and for predicting and stratifying risk? Do we look for higher-order effects only when a main genetic effect has been found? Do we try to fit a variety of models of interactions, including additive and multiplicative effects? Do we remain truly agnostic in our approach and let the data speak for themselves by using other approaches such as data mining techniques (24)? Do we continue using the multiplicative model to remove one dimension of complexity (25)? We need some analytic help to make the GEWIS efforts more productive by addressing biologic, clinical, and public health questions, not only academic abstractions!

EMERGING METHODS FOR ANALYSIS OF GEWIS

In this issue of the *Journal*, Murcay et al. (26) present a 2-step approach to evaluation of multiplicative gene-environment interaction in the context of a GEWIS. In another accompanying commentary, Chatterjee and Wacholder (27) discuss the strengths and limitations of this approach and compare it with a recently proposed (28) “1-stage” approach to gene-environment interaction. While analytic approaches to genomic data and gene-environment interaction will continue to evolve (refer to Chen et al. (29), Schwender and Ikstadt (30), Musani et al. (31), and Kraft et al. (32) for other examples), integrating analysis of genetic and environmental factors into a coherent biologic framework will be a huge challenge for epidemiology in the 21st century. Strangely enough, the current GWAS approach that took us away from biology and more toward the much-maligned “fishing expedition” in epidemiology provides more evidence about how much remains to be learned about the etiology of complex diseases. The traditional analytic tools that we have used in epidemiology have been strained by GWAS and will have to be further developed as we move from GWAS to GEWIS in the coming decades.

ACKNOWLEDGMENTS

Author affiliations: National Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, Georgia (Muin J. Khoury); and Division of Cancer

Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland (Sholom Wacholder).

The findings in this paper reflect the viewpoints of the authors and do not necessarily reflect the views of the Department of Health and Human Services.

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Conflict of interest: none declared.

REFERENCES

- Kennedy D. Breakthrough of the year [editorial]. *Science*. 2007;318(5858):1833.
- Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*. 2008;118(5):1590–1605.
- Collins FS, Green E, Guttmacher AE, et al. A vision for the future of genomics research. *Nature*. 2003;422(6934):835–847.
- International HapMap Consortium, Frazer KA, Ballinger DA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851–861.
- Rothman N, Wacholder S, Caporaso NE, et al. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta*. 2000;1471(2):C1–C10.
- Davey Smith G, Ebrahim S. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003;32(1):1–22.
- Pharoah PD, Antoniou AC, Easton DF, et al. Polygenes, risk prediction and targeted prevention of breast cancer. *N Engl J Med*. 2008;358(26):2796–2803.
- Hunter DJ, Altshuler D, Rader DJ. Focus on research: from Darwin’s finches to canaries in the coal mine—mining the genome for new biology. *N Engl J Med*. 2008;358(26):2760–2763.
- Kraft P, Hunter D. Integrating epidemiology and genetic associations: the challenge of gene-environment interaction. *Philos Trans R Soc Lond B Biol Sci*. 2005;360(1460):1609–1616.
- Seminara D, Khoury MJ, O’Brien TR, et al. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology*. 2007;18(1):1–8.
- Vineis P. Methodological approaches to gene-environment interactions in occupational epidemiology [electronic article]. *Occup Environ Med*. 2007;64:e3. (<http://oem.bmj.com/cgi/content/extract/64/12/e3>).
- Schwartz DA. The importance of gene-environment interactions and exposure assessment in understanding human diseases. *J Expo Sci Environ Epidemiol*. 2006;16(6):474–476.
- Yu W, Gwinn M, Clyne M, et al. A navigator for human genome epidemiology. *Nat Genet*. 2008;40(2):124–125.
- National Office of Public Health Genomics, Centers for Disease Control and Prevention. The Human Genome Epidemiology Network (HuGENet). (<http://www.cdc.gov/genomics/hugenet/default.htm>). (Accessed May 8, 2008).
- Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol*. 2002;156(4):300–310.

16. García-Closas M, Malats N, Silverman D, et al. NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*. 2005;366(9486):649–659.
17. Wacholder S, Chanock S, García-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004;96(6):434–442.
18. Khoury MJ, Adams MJ, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet*. 1988;42(1):89–95.
19. Khoury MJ, Beaty TH, Hwang SJ. Detection of genotype-environment interaction in case-control studies of birth defects: how big a sample size? *Teratology*. 1995;51(5):336–343.
20. Hunter DJ, Kraft P. Drinking from the fire hose. Statistical issues in genomewide association studies. *N Engl J Med*. 2007;357(5):436–439.
21. Hunter DJ. Gene-environment interactions in human disease. *Nat Rev Genet*. 2005;6(4):287–298.
22. Greenland S, Lash TL, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:71–83.
23. Hood L, Heath JR, Phelps ME, et al. Systems biology and new technologies enable predictive and preventative medicine. *Science*. 2004;306(5696):640–643.
24. Onkamo P, Toivonen H. A survey of data mining methods for linkage disequilibrium mapping. *Hum Genomics*. 2006;2(5):336–340.
25. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*. 2005;37(4):413–417.
26. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–226.
27. Chatterjee N, Wacholder S. Invited commentary: efficient testing of gene-environment interaction. *Am J Epidemiol*. 2009;169(2):231–233.
28. Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*. 2008;64(3):685–694.
29. Chen X, Liu CT, Zhang M, et al. A forest-based approach to identifying gene and gene-gene interactions. *Proc Natl Acad Sci U S A*. 2007;104(49):19199–198203.
30. Schwender H, Ikstadt K. Identification of SNP interaction using logic regression. *Biostatistics*. 2008;9:187–198.
31. Musani SK, Shriner D, Liu N, et al. Detection of gene \times gene interactions in genome wide association studies in human populations. *Hum Hered*. 2007;63(2):67–84.
32. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007;63(2):111–119.