Mapping complex disease loci in whole-genome association studies

Christopher S. Carlson[1], Michael A. Eberle[3], Leonid Kruglyak[2,3] & Deborah A. Nickerson[1]

[1]Department of Genome Sciences, University of Washington, 1705 NE Pacific, Seattle, Washington 98195-7730, USA (e-mail: csc47@u.washington.edu)
[2]Howard Hughes Medical Institute and [3]Division of Human Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA

Identification of the genetic polymorphisms that contribute to susceptibility for common diseases such as type 2 diabetes and schizophrenia will aid in the development of diagnostics and therapeutics. Previous studies have focused on the technique of genetic linkage, but new technologies and experimental resources make whole-genome association studies more feasible. Association studies of this type have good prospects for dissecting the genetics of common disease, but they currently face a number of challenges, including problems with multiple testing and study design, definition of intermediate phenotypes and interaction between polymorphisms.

After decades of study, much has been learned about the genetics of common diseases, but most of this knowledge relates to rare families segregating high-risk alleles[1]. Such alleles are generally very rare in the population and therefore explain relatively little of overall disease prevalence. A simple way of describing the relative importance of a locus from the standpoint of public health is to use the population attributable fraction (PAF). This factor can be thought of as the fraction of the disease that would be eliminated if the risk factor were removed. In a public health context, high-risk alleles may have a large PAF (>50%) in rare, single-gene mendelian diseases such as cystic fibrosis[2,3], but do not appear to be important for common diseases. The PAF for rare alleles in common disease is generally less than 10%. For example, more than 150 rare high-risk alleles have been identified for Alzheimer's disease in three genes (*PS1*, *PS2* and *APP*; for more information, see http://molgen-www.uia.ac.be/ADMutations), but the combined PAF for all of these alleles is less than 5% of all Alzheimer's cases[4].

Common modest-risk alleles may account for a greater PAF in common disease than do rare high-risk alleles; this is often referred to as the common disease/common variant (CDCV) hypothesis[5]. For example, the PAF for the Apoϵ4 allele in late-onset Alzheimer's disease has been estimated at 20% (ref. 6), which is greater than all of the known rare high-risk alleles combined.

Scientists are motivated by three factors in their search for common modest-risk variants associated with common disease: identified risk variants can illuminate new and important aspects of disease pathogenesis; common variants can be more important than rare ones from a public health perspective because they usually have a larger associated PAF; and common modest-risk variants are easier to identify than rare modest-risk variants[7], as will be discussed later. Association studies compare the allele frequency of a polymorphic marker, or a set of markers, in unrelated patients (cases) and healthy controls to identify markers that differ significantly between the two groups. Many papers have been published detailing studies using just one or a small number of polymorphisms, but many, if not most, of the identified associations have been difficult to replicate[8]. Because it seems likely that common modest-risk variants exist[8,9], developing tools such as high-density genetic maps for whole-genome association analyses is an important goal in the next phase of analysing the human genome. Comprehensive analysis of candidate regions, or even the whole genome, should produce more robust results. Over the past several years, the technical[10], informatic[11] and statistical[12] foundations have been laid for whole-genome association analyses[7,13–16]. It is important that researchers planning to use these resources understand that association analysis using dense maps, such as the HapMap[17], is fundamentally different from the family linkage studies used to identify rare high-risk alleles.

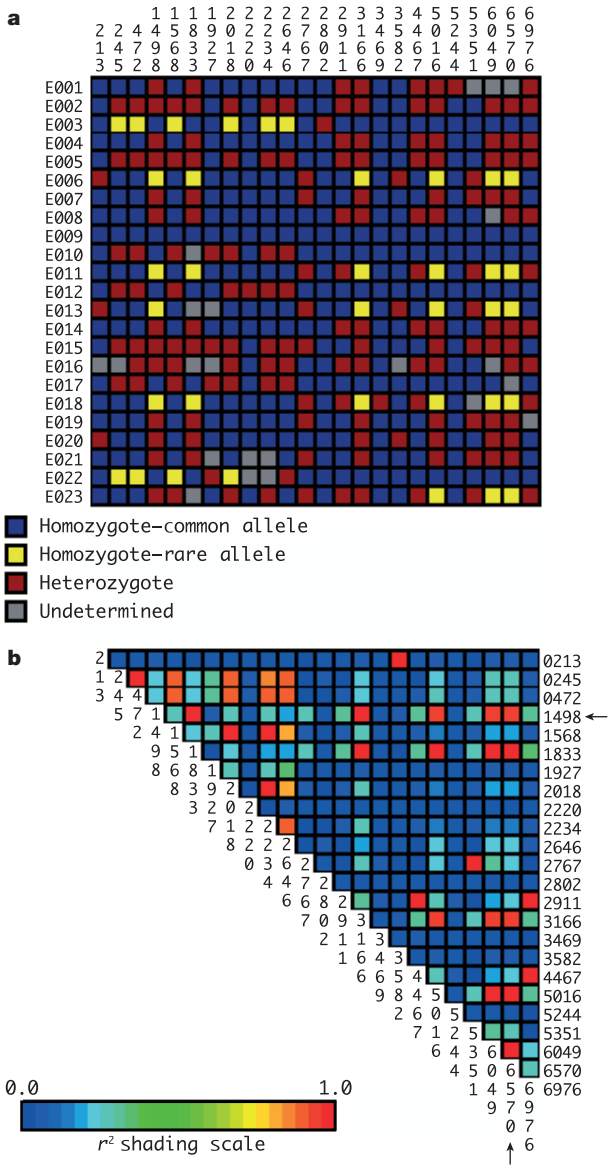
In this review, we will address the prospects for whole-genome association studies, starting with a summary of the theoretical basis of association analysis and currently available experimental resources. We will compare the strengths and weaknesses of direct, indirect and pooled study designs, and discuss the prospects for application of these designs at a whole-genome level. We close with a discussion of the challenges associated with multiple testing in whole-genome association analysis, both at the level of single polymorphisms and at the level of interaction between polymorphisms.

Linkage versus association

For the past two decades, the dominant study design for investigation of the genetic basis of inherited disease has been linkage analysis in families. Linkage analyses search for regions of the genome with a higher-than-expected number of shared alleles among affected individuals within a family. This indicates that somewhere within this linked region there is a disease-predisposing allele. Closely related individuals tend to share large regions of the genome inherited from the same recent ancestor, and therefore genotyping fewer than 500 polymorphic markers across the genome is generally adequate to detect linked regions. In linkage analysis, any polymorphism between a pair of linked markers will be associated with both markers. As a result, linkage maps have a logical hierarchical structure wherein an initial genome scan can be performed at low density (fewer than 500 markers), with additional markers used to fine-map the boundaries of linked regions. Because of the small number of recombination events within most families, it is frequently difficult to narrow the region of

Box 1

The non-hierarchical structure of linkage disequilibrium



In an indirect association analysis, a dense set of markers across a region is genotyped and tested for association. The assumption is that the genotype at polymorphisms in the region that are not genotyped directly will be correlated with one or more of the markers. The strength of the correlation between markers and an untyped polymorphism (frequently referred to as linkage disequilibrium, LD) can be described using several statistics. From the standpoint of association analysis, the most appropriate LD statistic is r^2, the Pearson correlation coefficient between alleles at each locus[25]. r^2 is directly related to statistical power to detect an untyped SNP. The ability to detect a risk SNP indirectly in n samples is roughly equivalent to the ability to detect it directly in nr^2 samples[61]. Thus, even in a large number of samples, low r^2 between markers and adjacent unassayed polymorphisms will result in low power to detect disease associations at the unassayed polymorphisms.

As an example, a visual genotype for the common SNPs in the *IL10* gene in a European population is shown in **a**, with r^2 between SNPs shown in **b** (image generated with the VG2 program; http://pga.gs.washington.edu/VG2.html). The important point is that even in a 7-kilobase, non-recombinant region (sometimes referred to as a 'haplotype block'), there are clearly several common patterns of variation, and levels of r^2 between these patterns are low. If we genotyped SNPs 1498 and 6570 as markers (indicated by arrows), we would observe strong LD between them ($r^2 = 1$; shown in red). But there are clearly a number of SNPs between these two with distinct patterns of genotypes (for example, 2767, 2911, 3582 and 4467) that are weakly associated with either marker and therefore would not be detected. Careful scrutiny will reveal that each common pattern is seen at least twice, effectively ruling out recurrent mutation as a source of variation in the region. This graphically illustrates how LD between adjacent markers that are not associated with a phenotype cannot be used to hierarchically exclude the intervening region from containing a disease-associated SNP.

interest to below several megabases, but candidate-gene analyses within linked regions have been greatly accelerated by the availability of the complete sequence of the human genome.

Linkage analysis is more powerful than association analysis for identifying rare high-risk disease alleles, but association analysis is expected to be more powerful for the detection of common disease alleles that confer modest disease risks[7]. This reflects the fact that for modest-risk alleles the patterns of allele sharing among affected individuals within pedigrees are less striking than patterns of allele sharing between unrelated affected individuals. Another advantage of association analysis is that it is easier to recruit large numbers of unrelated affected individuals than it is to collect large numbers of pedigrees, each containing multiple affected individuals, especially for diseases of old age. But the region around a marker that is shared identically by descent in unrelated, affected individuals will be much smaller than the shared region for related individuals because of the much higher number of generations from the most recent common ancestor. Therefore, association analysis requires markedly higher marker densities than linkage analysis, of the order of hundreds of thousands of markers[13,17].

Single-nucleotide polymorphisms (SNPs) are variations in DNA sequence where one of the four nucleotides is substituted for another (for example, C for A). SNPs are the most frequent type of polymorphism in the genome, and therefore will make up the vast majority of markers in a whole-genome association map. We will refer to SNPs genotyped in an association analysis as tagSNPs; a wide variety of techniques for selecting tagSNPs exist[10,18–23], with many more under development.

Linkage disequilibrium (LD) describes the non-random correlation between alleles at a pair of SNPs. A major practical difference between linkage maps and association maps is that LD does not have a hierarchical structure akin to linkage[12]. In association analysis, a SNP between a pair of strongly associated tagSNPs may not be strongly associated with either tagSNP (Box 1) and therefore can easily go undetected. Association analysis can identify disease-risk variants only when such variants are strongly associated with a tagSNP.

The first generation of whole-genome association maps will be available shortly, but as denser maps become available it will be necessary to genotype additional tagSNPs across the entire genome until

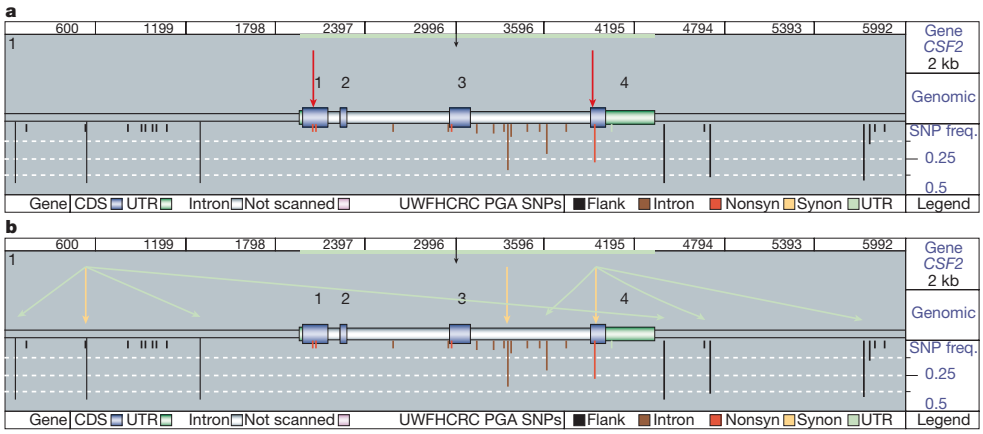


Figure 1 Direct versus indirect association analysis. **a**, In direct association analysis, all functional variants (red arrows) are catalogued and tested for association with disease. A GeneSNPs image of the *CSF2* gene is shown. Genomic features are shown as boxes along the horizontal axis (for example, blue boxes indicate exons). Polymorphisms are shown as vertical bars below the axis, with the length of the line indicating allele frequency and colour indicating context (for example, red indicates coding SNPs that change amino acids). **b**, For indirect association analysis, all common SNPs are tested for function by assaying a subset of tagSNPs in each gene (yellow arrows), such that all unassayed SNPs (green arrows) are correlated with one or more tagSNPs. Effects at unassayed SNPs (green arrows) would be detected through linkage disequilibrium with tagSNPs. Images adapted from GeneSNPs (http://www.genome.utah.edu/genesnps).

it becomes clear that most of the newly typed tagSNPs are strongly associated with tagSNPs that have previously been genotyped. Once patterns of LD between common polymorphisms have been described, it will be possible to select optimized subsets of tagSNPs such that all common variants (or a sufficiently large fraction thereof) are either directly assayed or are strongly correlated with either an allele of a single tagSNP or a combination of alleles at several tagSNPs (a tagSNP haplotype). The optimal subset will not necessarily be the same for all studies, but will reflect a number of practical considerations specific to the study, including the expected effect size and desired statistical power, the genetic diversity of the study population and the available budget. As genotyping and ultimately whole-genome sequencing becomes cheaper[24], map optimization may not be necessary at all.

Current applications of SNP maps

Linkage analysis generally identifies broad intervals of several megabases that correlate with disease status within pedigrees, and these intervals can encompass dozens to hundreds of candidate genes. If the disease-predisposing allele is relatively common, it should be possible to fine-map the disease allele within linkage peaks using LD[25]. Where a significant linkage hit has been identified and there is evidence of a single major-risk allele (as is found in haemochromatosis and cystic fibrosis), this is a perfectly reasonable strategy. However, a large number of 'suggestive' linkages exist for which the underlying genetic defect (if it exists) has not been identified. Suggestive linkage describes genomic regions with an observed trend toward excess sharing in affected individuals that is not significant after correcting for multiple tests across the genome[26,27]. Suggestive-linkage regions can contain common disease alleles with modest relative risk, so there is some enthusiasm for association analysis of candidate regions of the genome identified as suggestive linkages. These linkages can also be attributed to simple false-positive results, with one suggestive-linkage hit expected per linkage scan under the null hypothesis that no risk allele exists. Thus, fine-mapping of suggestive-linkage peaks by association analysis makes sense only when simulations show clear evidence for an excess of suggestive linkage across the genome[28], or when the cost of doing so is acceptable without compelling evidence for linkage.

Candidate SNP analysis of coding SNPs (cSNPs) that change amino acids has intriguing potential, because the number of common cSNPs is several orders of magnitude smaller than the number of common SNPs overall. In resequencing 262 genes in the PDR90 panel (which consists of 90 human DNA samples with diverse ethnic origins[29]), we have identified 147 cSNPs with an allele frequency above 5% (unpublished data). Extrapolating to 30,000 genes, a comprehensive analysis of common cSNPs could require ~20,000 common cSNPs to cover the genome, consistent with previous reports[30–32]. This approach is currently impractical because there is no concerted public effort to comprehensively identify cSNPs, but we believe that such a resource would be very valuable and should be made a high priority of the genome project.

cSNPs constitute the majority of disease alleles in mendelian disorders, and a recent analysis[1] suggests that common disease variants are likely to show a similar trend. Mendelian diseases are largely recessive, with mutations that abolish the normal function of a gene, as demonstrated by the relatively high frequency of nonsense mutations, so if common disease alleles reflect modest changes in gene function or activity, then the mendelian alleles are an inappropriate model. Furthermore, recent comparative genomic studies have demonstrated that the level of evolutionarily conserved noncoding sequence is comparable to the amount of evolutionarily conserved exonic sequence[33,34]. It seems quite plausible that disease-associated variants with modest effect will be distributed proportionately between noncoding and coding sequences. However, our ability to identify functional variation in conserved NCSs is still in its infancy[35,36].

Candidate SNP analysis is a direct test of association between a putatively functional variant and disease risk. The alternative, which we will refer to as indirect association (Fig. 1), is to test a dense map of SNPs for disease association under the assumption that if a risk polymorphism exists it will either be genotyped directly or be in strong LD with one of the genotyped tagSNPs[5,13,18]. The advantage of indirect association analysis is that it does not require prior determination of which SNP might be functionally important, but the disadvantage is that a much larger number of SNPs needs to be genotyped. Both direct and indirect association testing currently can be applied effectively to candidate genes that have been implicated in disease pathogenesis by other means, as long as common variants

have been comprehensively identified in the candidate gene, and the two approaches are not mutually exclusive.

Measured allele frequencies in pooled samples can be used in association analysis instead of genotyping individual samples[37–39]. Using quantitative genotyping technologies to measure allelic frequency changes between pools of patient and control DNAs is considerably cheaper than genotyping individual samples, even though multiple replicates are required to increase the precision of allele frequency estimates in pools. Because pooling is cheaper, many more SNPs can be screened for association in a pooled analysis. Pooled studies will no doubt find some associations[41–43], but will probably miss others for one of two reasons.

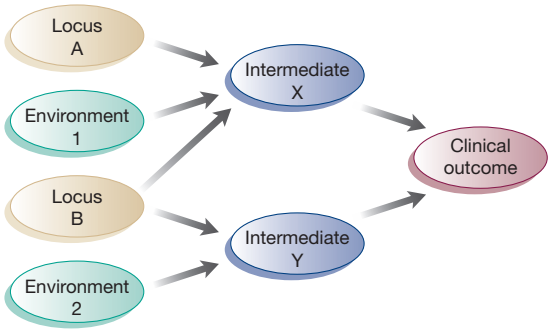
The power of pooled analysis depends on the expected effect size: large relative risks can lead to large differences in allele frequency (Δp) between cases and controls, whereas small relative risks (for example, at modest-risk alleles) produce smaller differences. The ability to detect a real allele frequency difference is therefore a function of the true Δp between cases and controls and the error in estimating Δp. Error in Δp estimates can be divided into sampling error (random noise introduced by estimating Δp in a finite sample) and measurement error (noise introduced by the precision of the technological platform used to estimate Δp). Whereas sampling error decreases with increasing sample size, measurement error does not, and seemingly small measurement errors in Δp can dramatically reduce the power to detect modest-risk alleles. This is especially true in whole-genome studies, where the large number of tests virtually guarantees that true positives will be lost in a sea of false positives. For example, the expected Δp is just 4.3% for a risk allele with 10% allele frequency and 1.5-fold genotype relative risk (GRR; the relative risk associated with each risk allele carried by an individual), which could easily be missed in a whole-genome scan with measurement error of $\pm 2\%$, as observed in most pooling platforms[44].

Technical considerations aside, pooled DNA analysis can be considerably less powerful than individual genotype analysis when known risk factors have been measured in each sample. This is true because common diseases generally have known environmental and genetic risk factors (for example, smoking status or sex). Analyses corrected for known risk factors should have greater power to detect novel factors when the risk associated with an unidentified genetic variant is smaller than that associated with a known factor, or if there is an interaction between the known risk factor and the unidentified genetic variant. This shortfall can be overcome by sorting cases and controls into subpools on the basis of known risk factors (such as smoking and sex, or genotype at a known risk locus) before the pools are genotyped. However, subpooling also increases the number of assays per SNP, so the economic advantage over individual genotyping is less marked. Even if subpools are assembled on the basis of known risk factors, the pooled frequency data cannot be used to explore novel gene–gene or gene–environment risk interactions.

The joy of intermediate phenotypes

Clinical outcome can be thought of as a synthesis (or grand total) of many risk factors, with intermediate phenotypes as subtotals. An example of this would be lipid profiles and risk of myocardial infarction, where the number of variables affecting low-density lipoprotein (LDL) levels is presumably smaller than the number of variables affecting the risk of myocardial infarction. Genetic factors contributing to intermediate phenotypes will generally be easier to identify because of the improved signal-to-noise ratio in the fraction of variance explained by any single factor (Box 2). Thus, it is crucial to appreciate how the definition of a phenotype can affect the prospects of an association analysis. Studies using a single clinical endpoint are akin to a shot at the moon with only one chance of success, whereas studies that collect multiple phenotypes are more likely to help us to understand the contribution of genetic factors to components of disease, regardless of whether a significant effect can be established on the clinical endpoint.

Box 2
Intermediate phenotype analysis



Intermediate clinical phenotypes have been identified for many diseases. Just as clinical outcome can be treated as a synthesis of the intermediate phenotypes, so intermediates can be treated as syntheses of subsets of proximal risk factors, both environmental (Environment 1 and 2) and genetic (Locus A and B), as shown in the figure. The advantage of intermediate phenotypes is that the number of genetic and environmental factors influencing each intermediate is presumably smaller than the number of factors affecting the clinical endpoint. Therefore, the proportion of variance in an intermediate phenotype explained by a given genetic locus will be greater than the proportion of variance explained in the clinical endpoint. A good example of this is the relationship between the gene *PON1* and risk of carotid artery disease (CAAD). The enzymatic activity of PON1 is a known risk factor in CAAD[62] and is a useful intermediate phenotype. Examination of genetic variants at the *PON1* locus has identified coding polymorphisms[63,64] and regulatory SNPs that significantly alter PON1 levels[65], but none of these SNPs showed significant associated risk for CAAD directly[62,66]. In this case, the intermediate phenotype (enzyme activity) helped in the identification of a group of SNPs that probably do have a small effect on CAAD risk, but a risk increase that is so small that it could not be detected directly in the available sample.

When intermediates are relatively cheap (for example, standard clinical chemistry) they should be broadly applied, although multiple testing can become an issue because each additional phenotype increases the number of statistical tests. More expensive intermediate phenotypes, such as RNA expression and protein levels, or measures of drug response and metabolism, also deserve thoughtful consideration, and new high-throughput proteomic approaches may dramatically expand the feasibility of collecting intermediate phenotype data[45]. Intermediate phenotypes may currently be a better investment than genotypes, at least until dense SNP-mapping technologies mature, especially considering that early studies will not necessarily help us to focus later efforts, because dense SNP maps do not have a hierarchical structure.

Although genotypes are reasonably precise, phenotypic and environmental risk factors are generally imprecisely measured, and the precision of these covariates will be very important in evaluating gene–environment interactions. For example, body mass index (BMI) is a more precise (and useful) measure than is a dichotomous obese/non-obese variable. For that matter, circulating levels of the nicotine metabolite cotinine are certainly a better measure of smoking status than self report[46]. Just as intermediate phenotypes increase the probability of learning something useful from a study, a modest number of carefully phenotyped samples can be more valuable than a large number of poorly characterized samples. Conversely, studies that make a reasonable effort to regulate environmental exposures

should be better able to identify modest genetic effects because of reduced environmental noise.

Can the multiple testing problem be solved?

Before dense SNP maps are applied to search for genetic components of a complex disease, we need to understand how many markers will be required for a complete, high-power genome scan. More than seven million SNPs have been described in the human genome. The International HapMap Project[17] aims to genotype at least 600,000 of these to define haplotype patterns across the genome, with the goal of developing a map of non-redundant tagSNPs, and is committed to denser spacing where necessary. Ultimately, such a map is likely to require 200,000 to one million markers to achieve a reasonable likelihood that any common SNP in the genome is usefully associated with at least one tagSNP[17], which is consistent with previous theoretical estimates[13]. This generates an enormous multiple testing problem, and it will be difficult to achieve 'statistical significance' in a single genome-wide association study for all but the strongest effects. However, there are practical methods that may help to identify important genetic factors more efficiently, such as ranking markers by proximity to candidate genes or by expected functional consequence (for example, cSNPs).

Previous analyses have estimated the ability to detect a given GRR in the context of a case–control study. In one model of a genome-wide association study, a GRR of >1.5 associated with a common SNP (minor allele frequency of 10%) could be detected in a sample of ~2,000 matched case–control pairs[7], conservatively corrected for one million independent tests. More recent analysis suggests that detecting a similar GRR would require many more samples, but for this the investigators considered a map of only 300,000 tagSNPs[47]. These studies raise an important issue: even if it is difficult to establish significance within a single study, initial scans will at least help to focus subsequent efforts on interesting regions of the genome. Furthermore, these analyses conservatively assumed independent markers, but many markers in SNP maps will be significantly associated with one another. Thus, it is difficult to theoretically establish a threshold for significance in whole-genome association analysis, and significance will be easier to address empirically by permutation analysis of the observed data. False-discovery-rate analysis may also prove to be a useful tool in this area[48].

Higher-order risk interactions

One of the possible reasons that linkage analysis has identified few high-frequency, modest-risk alleles is statistical interaction (or epistasis) between multiple loci. If the risk associated with a given locus is influenced by the genotype at another locus, single-locus analyses might miss the association. Given that a large proportion of common disease risk is heritable and that the genetic risk factors identified by linkage explain relatively little of the PAF for common disease, epistatic interactions could well be important in predicting disease risk. If multiple testing is a major challenge at the level of single SNPs, the problem rapidly becomes intractable when we allow for gene–gene or gene–environment interactions. That is, for 600,000 tagSNPs and ten environmental variables, there are more than 10^{11} possible pairwise gene–gene interactions, and six million possible gene–environment interactions. Can we logically limit the search space for gene–gene and gene–environment interactions?

Overall disease risk can be modelled as the product of risks at many independent risk loci. With such a model, high-risk combinations of genotype will exist, but the ability to detect any single locus is a function of the relative risk of that locus alone. So what exactly is an interaction? Statistical interaction between loci requires a dependent effect, wherein the risk associated with a genotype at a locus is dependent on a genotype at another locus. There are several possible epistatic models. Assuming two risk loci (A and B), a synergistic epistatic effect occurs when the combined risk to risk-allele carriers at both loci is greater than would be expected if they were independent, and an antagonistic epistatic effect is when the combined risk is lower than expected. A balanced epistatic effect is a special case of antagonistic effects where the risk for risk-allele carriers at both loci is actually lower than for risk-allele carriers at only one locus. Finally, a permissive epistatic interaction is when only risk-allele carriers at both loci show increased risk of disease.

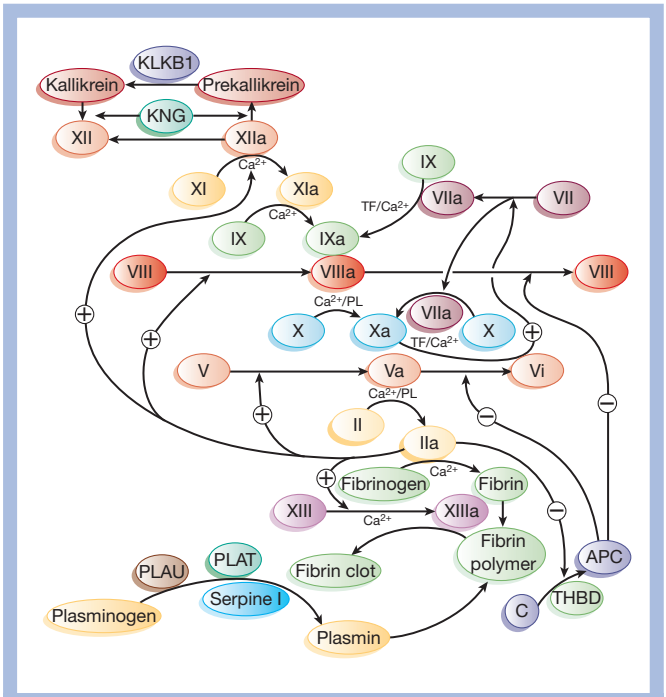


Figure 2 Pathway of physical interactions. If the risk associated with a given locus is influenced by another locus, single-locus effects may not be detected at either locus. Identifying interacting loci has improved with the introduction of expertly curated pathways like the one shown here for genes involved in blood clotting. Each gene is represented as an oval, with plain arrows representing conversion of proteins from one form to another (for example, VII to VIIa). Arrows with + or − symbols indicate enzymes that facilitate or inhibit conversions (for example, Xa converts VII to VIIa). Each type of epistatic risk can conceivably take place in such a pathway. For example, balanced risk interactions are plausible for physically interacting proteins (for example, factor IIa/fibrinogen), whereas permissive interactions are plausible for genes that are involved in the same pathway stream (for example, factors VIIa/II).

Although relatively few epistatic interactions have been described to date[49–52], there is little doubt that many such interactions exist. But the sheer number of potential interactions practically guarantees that a comprehensive search has no power to detect them. In practice, there are several reasonable approaches to reduce the number of interactions analysed, such as limiting analysis to biologically plausible interactions between genes in related pathways, or limiting analysis to markers with appreciable single-locus effects. In general, risk interactions are more plausible between genes involved in a physical interaction, found in the same pathway or involved in the same regulatory network. Otherwise it is difficult to construct a model in which risk at one locus can be dependent on the genotype at the second locus. It would be reasonable to test for epistatic interactions between all pairs of non-synonymous cSNPs in physically interacting genes or pathways. However, identifying interacting genes requires improved bioinformatics tools to expertly annotate known pathways of physical interaction[53,54] (Fig. 2), mine gene-ontology resources[55,56] and annotate expression analyses for evidence of co-regulation[57]. Fortunately, all of these are likely to become a reality in the near future.

Another approach to limiting the search space is to consider only interactions between SNPs that exhibit some evidence of a main or

single-locus effect. Epistatic risks with a large PAF are likely to exhibit large main effects, so it might be reasonable initially to restrict epistatic analysis to SNPs with large observed main effects. A test statistic conditioned on the observed marginal genotype frequencies at each SNP in cases could be used to identify epistatic interactions within cases, independently from the epistasis model or main effect at either locus. Similar tests for interaction have been described for gene–environment interactions, where the magnitude of an interactive effect can accurately be estimated from case data alone[58–60]. Testing the 1,000 tagSNPs with the largest main effects for pairwise epistatic interaction would result in ~500,000 tests, comparable to the number of single-locus tests in a whole-genome scan. Conveniently, significant epistatic interactions among these tagSNPs can provide independent evidence that the main effects are real. As a final note on the exploration of epistatic interaction, the methodologies suggested above are solely an attempt to maximize power to identify statistically significant results within a single study. After the statistical power of a given study has been exhausted, it is still important to examine all possible interactions, although such a search is computationally daunting. Complete exploration of non-obvious interactions can help refine underlying analytical assumptions and study design, as well as generate testable hypotheses for subsequent efforts.

Future directions

The tools for the genetic dissection of complex traits are nearing maturity, and dense SNP maps will allow researchers to test the CDCV hypothesis and identify such variants where they exist. These discoveries will complement the knowledge already gleaned from rare mendelian disorders, providing information about which gene pathways are important in disease pathogenesis, which genes and alleles within these pathways are important risks in the general population, and how much of the heritable risk for common disease remains to be explained by rare modest-risk alleles and epistatic interactions. These tools may also facilitate the identification of important risk interactions, but considerable theoretical efforts to develop test statistics will be necessary before this can be achieved. Common variants identified will not wholly explain the genetic component of common disease risk, as there are certain to be rare modest-risk variants that will go undetected by these analyses, but the identification of important common variants will provide new targets for diagnostics, prognostics and therapeutics. □

doi:10.1038/nature02623

1. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (suppl.), 228–237 (2003).
2. Grody, W. W. *et al.* PCR-based screening for cystic fibrosis carrier mutations in an ethnically diverse pregnant population. *Am. J. Hum. Genet.* **60,** 935–947 (1997).
3. Kosorok, M. R., Wei, W. H. & Farrell, P. M. The incidence of cystic fibrosis. *Stat. Med.* **15,** 449–462 (1996).
4. Rocchi, A., Pellegrini, S., Siciliano, G. & Murri, L. Causative and susceptibility genes for Alzheimer's disease: a review. *Brain Res. Bull.* **61,** 1–24 (2003).
5. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278,** 1580–1581 (1997).
6. Slooter, A. J. *et al.* Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: the Rotterdam Study. *Arch. Neurol.* **55,** 964–968 (1998).
7. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
8. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33,** 177–182 (2003).
9. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510 (2001).
10. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–1723 (2001).
11. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
12. Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5,** 89–100 (2004).
13. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22,** 139–144 (1999).
14. Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* **71,** 1386–1394 (2002).
15. Risch, N. & Teng, J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8,** 1273–1288 (1998).
16. Risch, N. Evolving methods in genetic epidemiology. II. Genetic linkage from an epidemiologic perspective. *Epidemiol. Rev.* **19,** 24–32 (1997).
17. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
18. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74,** 106–120 (2004).
19. Ke, X. & Cardon, L. R. Efficient selective screening of haplotype tag SNPs. *Bioinformatics* **19,** 287–288 (2003).
20. Stram, D. O. *et al.* Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.* **55,** 27–36 (2003).
21. Weale, M. E. *et al.* Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.* **73,** 551–565 (2003).
22. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29,** 233–237 (2001).
23. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
24. Collins, F. S., Green, E. D., Guttmacher, A. E. & Guyer, M. S. A vision for the future of genomics research. *Nature* **422,** 835–847 (2003).
25. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29,** 311–322 (1995).
26. Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11,** 241–247 (1995).
27. Sawcer, S. *et al.* Empirical genomewide significance levels established by whole genome simulations. *Genet. Epidemiol.* **14,** 223–229 (1997).
28. A meta-analysis of whole genome linkage screens in multiple sclerosis. *J. Neuroimmunol.* **143,** 39–46 (2003).
29. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8,** 1229–1231 (1998).
30. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22,** 239–247 (1999).
31. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22,** 231–238 (1999).
32. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293,** 489–493 (2001).
33. Schwartz, S. *et al.* MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31,** 3518–3524 (2003).
34. Mayor, C. *et al.* VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047 (2000).
35. Boffelli, D. *et al.* Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299,** 1391–1394 (2003).
36. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
37. Bansal, A. *et al.* Association testing by DNA pooling: an effective initial screen. *Proc. Natl Acad. Sci. USA* **99,** 16871–16874 (2002).
38. Mohlke, K. L. *et al.* High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl Acad. Sci. USA* **99,** 16928–16933 (2002).
39. Howell, W. M., Evans, P. R., Wilson, P. J., Cawley, M. I. & Smith, J. L. HLA class II DR, DQ, and DP restriction fragment length polymorphisms in rheumatoid arthritis. *Ann. Rheum. Dis.* **48,** 295–301 (1989).
40. Yang, Y. *et al.* Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl Acad. Sci. USA* **100,** 7225–7230 (2003).
41. Permutt, M. A. *et al.* Searching for type 2 diabetes genes on chromosome 20. *Diabetes* **51** (suppl. 3), S308–S315 (2002).
42. Barcellos, L. F. & Thomson, G. Genetic analysis of multiple sclerosis in Europeans. *J. Neuroimmunol.* **143,** 1–6 (2003).
43. Kammerer, S. *et al.* Amino acid variant in the kinase binding domain of dual-specific A kinase-anchoring protein 2: a disease susceptibility polymorphism. *Proc. Natl Acad. Sci. USA* **100,** 4066–4071 (2003).
44. Sham, P., Bader, J. S., Craig, I., O'Donovan, M. & Owen, M. DNA pooling: A tool for large-scale association studies. *Nature Rev. Genet.* **3,** 862–871 (2002).
45. Cristea, I. M., Gaskell, S. J. & Whetton, A. D. Proteomics techniques and their application to hematology. *Blood* **103,** 3624–3634 (2004).
46. Haufroid, V. & Lison, D. Urinary cotinine as a tobacco-smoke exposure index: a minireview. *Int. Arch. Occup. Environ. Health* **71,** 162–168 (1998).
47. Schork, N. J. Power calculations for genetic association studies using estimated probability distributions. *Am. J. Hum. Genet.* **70,** 1480–1489 (2002).
48. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100,** 9440–9445 (2003).
49. Bolk, S. *et al.* A human model for multigenic inheritance: phenotypic expression in Hirschsprung disease requires both the *RET* gene and a new 9q31 locus. *Proc. Natl Acad. Sci. USA* **97,** 268–273 (2000).
50. Zetterberg, H., Zafiropoulos, A., Spandidos, D. A., Rymo, L. & Blennow, K. Gene–gene interaction between fetal MTHFR 677C>T and transcobalamin 776C>G polymorphisms in human spontaneous abortion. *Hum. Reprod.* **18,** 1948–1950 (2003).
51. Butt, C. *et al.* Combined carrier status of prothrombin 20210A and factor XIII-A Leu34 alleles as a strong risk factor for myocardial infarction: evidence of a gene–gene interaction. *Blood* **101,** 3037–3041 (2003).
52. Tiret, L. *et al.* Synergistic effects of angiotensin-converting enzyme and angiotensin-II type 1 receptor gene polymorphisms on risk of myocardial infarction. *Lancet* **344,** 910–913 (1994).
53. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet.* **31,** 19–20 (2002).

54. Bonner, A. E., Lemon, W. J. & You, M. Gene expression signatures identify novel regulatory pathways during murine lung development: implications for lung tumorigenesis. *J. Med. Genet.* **40,** 408–417 (2003).
55. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32,** D258–D261 (2004).
56. Zeeberg, B. R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
57. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–210 (2002).
58. Khoury, M. J. & Flanders, W. D. Nontraditional epidemiologic approaches in the analysis of gene–environment interaction: case–control studies with no controls! *Am. J. Epidemiol.* **144,** 207–213 (1996).
59. Begg, C. B. & Zhang, Z. F. Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol. Biomarkers Prev.* **3,** 173–175 (1994).
60. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Stat. Med.* **13,** 153–162 (1994).
61. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69,** 1–14 (2001).
62. Jarvik, G. P. *et al.* Paraoxonase (PON1) phenotype is a better predictor of vascular disease than is PON1(192) or PON1(55) genotype. *Arterioscler. Thromb. Vasc. Biol.* **20,** 2441–2447 (2000).
63. Adkins, S., Gan, K. N., Mody, M. & La Du, B. N. Molecular basis for the polymorphic forms of human serum paraoxonase/arylesterase: glutamine or arginine at position 191, for the respective A or B allozymes. *Am. J. Hum. Genet.* **52,** 598–608 (1993).
64. Humbert, R. *et al.* The molecular basis of the human serum paraoxonase activity polymorphism. *Nature Genet.* **3,** 73–76 (1993).
65. Brophy, V. H. *et al.* Effects of 5′ regulatory-region polymorphisms on paraoxonase-gene (*PON1*) expression. *Am. J. Hum. Genet.* **68,** 1428–1436 (2001).
66. Jarvik, G. P. *et al.* Paraoxonase activity, but not haplotype utilizing the linkage disequilibrium structure, predicts vascular disease. *Arterioscler. Thromb. Vasc. Biol.* **23,** 1465–1471 (2003).

Acknowledgements This work was supported by grants from the National Heart, Lung and Blood Institute, the National Institute of Environmental Health Sciences and the National Institute of Mental Health. L.K. is a James S. McDonnell Centennial Fellow. Thanks to D. Altshuler for helpful input, to G. Jarvik, P. Heagerty and P. Scheet for discussions on epistatic risk models, and to T. Banghale, D. Crawford, B. Livingston, R. Mackelprang and M. Rieder for comments on the manuscript.

Competing interests statement The authors declare competing financial interests: details accompany the paper on **www.nature.com/nature**.