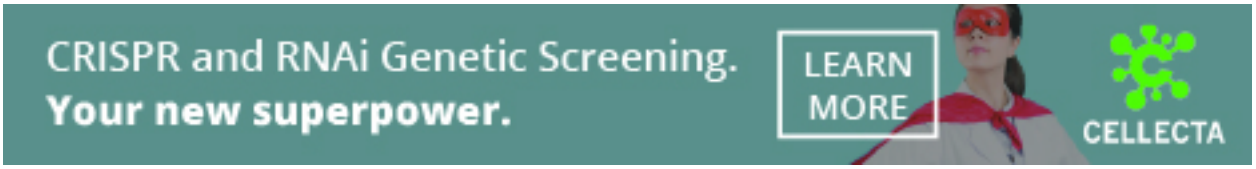
Determinants of the success of whole-genome association testing

Andrew G. Clark, Eric Boerwinkle, James Hixson, et al.

Genome Res. 2005 15: 1463-1467
Access the most recent version at doi:10.1101/gr.4244005



Perspective

Determinants of the success of whole-genome association testing

Andrew G. Clark,[1,4] Eric Boerwinkle,[2] James Hixson,[2] and Charles F. Sing[3]

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA; [2]Human Genetics Center, University of Texas Health Science Center, Houston, Texas 77030, USA; [3]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA

The convergence of an aging population and spiraling health care costs have produced a perfect storm of urgency for understanding the causal basis for common chronic disorders such as diabetes and cardiovascular disease. At issue is the role of genetics in providing solid, practical answers to contemporary public health challenges. The International HapMap Project has staked out ambitious claims, and the time to deliver on these promises approaches. We anticipate that analyses of HapMap data will suggest better study designs for identifying genetic variants that predict risk of chronic complex disorders. These studies will entail enormous costs, yet the answers to many important questions about the optimal design remain unanswered. The brave and the wealthy will be able to proceed without answers to these questions, and, indeed, some have experienced success (Hinds et al. 2004; Klein et al. 2005). At a time when many research groups are about to leap into the enterprise of whole-genome association testing, it seems prudent to reflect on assumptions that are being made and challenges that lie ahead. Our intent is not to question the fundamental idea of linkage disequilibrium mapping (Weiss and Terwilliger 2000), but rather to point out the hidden and untested assumptions implicit in using tag SNPs for linkage disequilibrium mapping of common diseases using samples from human populations. Our hope is to avoid expensive mistakes that may emerge if we ignore the assumptions related to study design and analysis, and blindly generate genotypes for large population studies with the hope that somehow LD mapping analysis will produce useful results.

Some of the factors that will have an impact on the efficacy of HapMap for providing guidelines for whole-genome association testing can be summarized as follows:

1. What is the population of inference?

The first design issue in any study of complex disorders is to identify and characterize the population about which inferences are sought. Most investigators ignore this problem and simply work with convenient samples. But it is critical to think carefully about the sampling design, and what population that sample is supposed to represent. The sampling design will determine whether inferences will apply only to that population, and how the conclusions will be used to inform public health decisions that disregard heterogeneity of genome–phenotype relationships among populations. One might think that finding genes that predict a disease that has a complex multifactorial etiology is like any other reductionist molecular biology problem, except that now our laboratory happens to include free-living human populations. This is a fallacy that needs to be dispelled. Population-level problems require population-level thinking from the outset. Complex diseases cannot be understood in terms of a deterministic clockwork universe; but, instead, they are inherently and intrinsically characterized by a tangle of probability densities that describe the complex networks of interacting genetic and environmental causal agents that are embedded in the high-dimensional pathways connecting the genome with disease-related phenotypes.

At the root of disease causation there are certain to be molecular mechanisms, but there is an enormous gulf between the detection of a disease association by linkage disequilibrium and the understanding of disease mechanisms. One synergistic intersection is the identification and study of candidate genes. Candidate genes provide an entry way to unraveling genetic causation, and success with model organisms such as *Drosophila* and mouse provide some hope for human candidate gene studies (Zwick et al. 2000). The field of epidemiology seeks to identify disease risk factors by collecting large and diverse samples and keeping track of many ancillary variables so that agents that may be involved in causation can be sorted out by subsequent stratification of the sample. But this method too has severe limitations, and can at best only deal with three or four variables at a time. At least it avoids the trap of limiting inferences to one context. An additional concern relating to the population of inference is whether there is hidden stratification in the sample, an attribute that can easily produce false positives (Hirschhorn and Daly 2005).

2. Efficacy of tag SNPs

If a pair of SNPs displays a high degree of linkage disequilibrium, then by obtaining information on only one of them, it is possible to predict with some confidence the genotype of the unobserved SNP. Essentially the tag SNP idea is equivalent to prediction of missing genotypic data. If we score genotypes of some (tag) SNPs, will we be able to predict the genotypes at unobserved sites? The entire HapMap project was launched in anticipation of a positive answer to this question, and now that the data are in, there are extensive efforts to optimize the use of the HapMap genotype data for identifying informative subsets of SNPs. The ability to make useful predictions in this way depends on the population of inference, and whether the allele frequencies and patterns of linkage disequilibrium in that particular population are accurately mirrored by the HapMap study populations and the relatively small number ($n = 90$) of individuals sampled from those populations (Zondervan and Cardon 2004). For a restricted sampling of human populations, the HapMap and ENCODE data will give us a reasonably good quantitative assessment of the ability to predict most other SNPs within those same populations, but

[4]Corresponding author.
E-mail ac347@cornell.edu; fax (607) 255-6249.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4244005. Freely available online through the *Genome Research* Immediate Open Access option.

extrapolation outside these populations will be much more tenuous. Although there are plans to expand HapMap genotyping to include additional populations, the utility of tag SNPs outside of the reference populations that are being considered remains untested. Particularly vexing is the ability of the Yoruban population sample from Nigeria to identify tag SNPs that will be informative about African American populations. Data on the patterns of linkage disequilibrium suggest rather large discrepancies between these two populations (Sawyer et al. 2005). The test of tag SNPs so far has rested on the ability to predict masked SNPs from the same data. An important assumption is that the disease-risk-inflating SNPs have the same frequencies and LD structure as the SNPs in HapMap samples. If they are on average rarer, or for some reason display less LD, the efficacy of tag SNPs could be overestimated.

3. The relative frequency of disease alleles

The efficacy of association mapping depends on the genetic architecture of the disease. All else being equal, if the primary genetic determinants of disease risk are rare alleles, they will be much more challenging to identify than will common variants (Wang et al. 2005). With limited power, a variant whose relative frequency is on the order of 0.01 can be identified by LD with common tag SNPs, but it will require sample sizes in the thousands (Zondervan and Cardon 2004). Variants rarer than 1% cannot be mapped effectively by LD. But even common variants may be challenging to detect if their role in disease risk is mediated through complex interactions or if their effects are small (see Factor #4 below). If combinations of common variants from multiple loci define rare genotypes with appreciable disease risk, then approaches that have been discussed for genome-wide association testing will miss these rare genotypes, even if they result in high disease susceptibility. Recent efforts have provided evidence that rare variants in established candidate genes might contribute to disease risk to a greater degree than originally appreciated (Cohen et al. 2004; Frikke-Schmidt et al. 2004). For genes having an obvious and pivotal role in a disorder, one can sequence the gene in individuals with the most extreme phenotypes and ask whether the distribution of rare variants, or rare multilocus genotypes, reflects a departure from what is expected if they had no effect on phenotype. Although such methods do not identify the role of any particular rare variant, they provide support for the role of rare variants in contributing to disease risk. In the end, what will likely emerge is that the genetic architecture of common diseases consists of gene variations with a spectrum of relative allele frequencies, and the shape of the distribution of these frequencies will likely differ across diseases.

4. The magnitude of allelic effects

Alleles with large effects on disease risk are more easily affected by natural selection, and thus they are likely to be maintained at low frequency owing to mutation–selection balance, unless they had been favorable in some previous environment. Classical Mendelian disorders whose alleles have large effects may attain high frequencies in some populations, and the textbook accounts of β^S-globin and malaria, Duffy null and *Plasmodium vivax*, and CCR5 Δ32 and HIV get sufficient attention that we tend to think that any disease allele at elevated frequency must have been favorable in the past. In reality, we have no idea how common this scenario is, given the strong ascertainment bias for finding alleles whose expressivity is high. On the other hand, given the amount of effort already expended to test associations between SNPs and, say, schizophrenia, it seems unlikely that there will be many remaining common variants whose increment on this disease risk is appreciable. All else being equal, common alleles will more likely have only a small effect on risk. This is because the larger the effect, the less likely it is to be maintained in a polymorphism, and there are so many genetic ways to produce a small effect. Many of the association studies now underway are large enough that alleles that explain only 1% of the variance in risk will likely be detectable as statistically significant. The problem may then turn into a policy issue—what should the public health policy be when considering alleles of such small effect? At what magnitude of effect is it worth doing diagnostic testing? Perhaps a composite genetic risk score that combines the effects of multiple loci will prove to be of more practical utility for common chronic diseases.

5. The independence of genetic effects

A fundamental issue in the analysis of SNP data is to define the unit of genetic function that influences disease risk. Is it a single SNP, a regulatory motif, an encoded protein subunit, a combination of SNPs in a combination of genes, an interacting protein complex, or a metabolic or physiological pathway? We almost never know the answer initially, and thus a proper study must allow for flexibility in interpretation of the unit of gene function. In model organisms, it seems that whenever a study is designed to detect gene–gene interactions, they are nearly always present (Anholt et al. 2003; Peripato et al. 2004; Brem and Kruglyak 2005). Because the statistical power to detect interaction effects is generally lower than the ability to detect main effects, it is remarkable that there are many cases in which interactions are detected in the absence of marginal effects (Hamon et al. 2004). There are sound reasons to expect that biology works through gene interactions, but the problem explodes when interactions are considered, because for every n SNPs or genes there are $n(n-1)/2$ pairwise interactions. A genome-wide scan with 1 million SNPs (3 kb coverage) will afford 10^{12} possible pairwise tests of SNP by SNP interactions. If these SNPs could be clustered into 30,000 genes, then the same data may identify four orders of magnitude fewer gene–gene interactions. In either case, statistical inference becomes a serious challenge with numbers like these. The preceding has discussed the issue of the functional genetic units from a biological/etiological perspective. Equally important is to define the prediction unit—what consideration of SNPs, haplotypes, or multigene complexes is most informative for statistically predicting disease risk? Because SNP data can be collected in a way that allows multiple models of prediction to be tested, we will have many opportunities to perform tests of the utility of different prediction units over the next few years.

6. Genotype by environment interaction

There is little argument that common diseases are the result of both genetic and environmental susceptibilities and their lifelong interactions. However, there is a paucity of research incorporating genotype-by-environment interaction into human genetic studies. Despite a plethora of editorials and perspectives that expound on its importance, most studies make an early and prominent assumption that no genotype-by-environment interactions exist. In fact, one explanation for common disease research falling short of earlier expectations is the inability to in-

corporate genotype-by-environment interactions in study designs and statistical analyses. There are likely multiple reasons for a shortage of genotype-by-environment interaction studies. The first is a consequence of difficulties in measuring individual exposures to the environment. Although the past decades have seen revolutionary advances in measurement of genetic variation, measurement of environmental factors is still largely based on questionnaires and indirect assessments. The second reason is fear of further exponential explosions in the number of variables with which we have to contend (see Factor #8 below). It is fair to suggest that if genetic susceptibility to the common chronic diseases is the result of gene-by-gene-by-environment interactions, the biological details of these interactions may be cloaked in uncertainty for years to come, and their utility for statistical modeling of disease risk may never be realized. The third reason is a lack of appreciation for good study design for estimating genotype-by-environment interaction. Arguably, the best study designs are those based on environmental modifications, such as changes in diet or cigarette smoking. However, such studies are expensive and difficult to execute in large samples. The frequency of large clinical trials and the ability to add a genetic component to these studies may be one reason for the increased popularity and success of pharmacogenetics, the study of a particular kind of genotype-by-environment interaction. A fourth reason for our inability to directly incorporate environmental factors into our genetic studies is cultural. Most investigators who are familiar with state-of-the-art genotyping technologies and the utility of the HapMap results are likely to consider the results of food frequency and exercise questionnaires to be soft or uninformative, and not worthy of further serious consideration. It should be obvious to most that if the applications of HapMap results are to attain their potential promise, the barriers to relaxing the assumption of no genotype-by-environment interaction need immediate attention.

7. The central role of longitudinal analysis

Genetic variation is expected to influence the initiation, progression, onset, and severity of a common chronic disease. Phenotype–genotype relationships are dynamic over the course of life (Sing et al. 2004). Initiation of the development of disease begins long before the phenotype reaches the clinical horizon. At a particular point in time in the life cycle, individuals with a particular genotype have a range of possible phenotypes determined by the range of possible environmental histories. The interaction of genotypes with time- and space-dependent exposures to environmental agents means that many individuals who have a particular genotype that is associated with an increased risk of disease may remain healthy because of exposures to compensatory environments. The converse will also be true. Few genetic studies take into account the reality of the dynamic relationships between an individual's genotype and history of exposure to environmental agents in predicting phenotypic outcomes at some future point in time in a particular environmental niche. Meaningful longitudinal studies are far more difficult to design and manage over the course of the decades required to carry them out. Measurement of the history of exposures to environmental agents in all genetic studies must be given a higher priority. Careful thought and organizational wisdom will be required to collect data that are representative of the population of inference. Measures of the contemporary environment may have diminished value compared to longitudinal measures in evaluating the role of genotype-by-environment interaction in predicting future outcomes.

8. Dimension catastrophe

There are two aspects of the problem of high dimensionality in making inferences about relationships between SNPs and disease risk. The problem is there are simply so many tests that methods for reducing this number to a manageable level are crucial. With an explosion in the numbers of tests being done, the power of the tests plummet and the number of false positive tests gets to be unmanageable. Approaches to dimension reduction, including machine learning methods, have made great strides in the recent past, but this problem is far from being solved. The second aspect of the dimension catastrophe comes from inferring causation when multiple factors are at play. If an important factor (genetic or environmental) is not considered, collapsing of the data along that factor may induce spurious interactions among the remaining variables. A model that ignores context can lead us badly astray, either producing spurious interactions or erasing interactions that exist if the context effects are accounted for. Many epidemiologists and statisticians are aware of the "reversal paradox," also known as Simpson's paradox, when there is a qualitative reversal in the direction of an effect across strata (Tu et al. 2005; Weinberg 2005). This is most readily seen by collapsing multidimensional contingency tables, where even simple χ^2 statistics can produce spurious interactions of marginal effects, or can make interactions disappear when one or more dimensions are collapsed (Simpson 1951). The challenge is that we cannot know up front which genetic or environmental strata are crucial, and thus we have to make guesses as to how to begin the dimension reduction, or, alternatively, apply methods of association testing that are robust to large numbers of independent variables. One needs to know the genetic unit to study gene-by-gene and gene-by-environment interactions, but we also need to accommodate the high dimensionality of genetic and environmental variables, taking into account stratifications that define the population of inference in order to define the genomic unit of inference.

9. Heterogeneity

Most complex disorders have multiple causes, so that a collection of cases may have considerable genetic heterogeneity, with disjoint causal mechanisms at play. The ability to deal with such heterogeneity in genetic and environmental causes, and in age-of-onset effects, will become crucial to success. It seems likely that some small number of cases will be associated with allelic variations of intermediate effect at single loci, but that the bulk will instead be associated with particular combinations of common alleles at many loci. The problem is to determine which combination of SNPs predicts risk in which subset of individuals. Rare single SNP and single gene variations cannot explain an appreciable amount of the total population risk. Enough studies have already been done to realize that the marginal effects of most SNPs are likely to be close to zero. It has been pointed out time and time again that single gene variations explain <2% of risk of a complex discrete endpoint like cancer or cardiovascular disease. If each SNP explains <2%–3% of the risk, then, because the genetic component for many complex disorders explains about half the variance in risk, it seems inescapable that there must be 20 or more different combinations of SNPs needed to

explain the full genetic component of risk, and that these factors will be heterogeneous among patients.

10. Validation

Even after carrying out appropriate dimension-reduction steps and adjusting for multiple comparisons as part of the statistical analysis, studies will need to be repeated in other populations to ask whether the results of a particular study have general applicability. Replication has become the sine quo non expectation for large-scale genomic association studies. Lack of replication has been an all too familiar characteristic of genetic association studies (Hirschhorn and Altshuler 2002). But in studies of human populations, there really are no replicate samples. The seemingly simple concept of replication carries with it a set of assumptions and challenges that have not been satisfactorily addressed. First, failure to repeat a result in a second sample may be the result of a type II error (i.e., false-negative result) and/or inadequate statistical power in the second (or third or fourth) sample. Second, because human populations are not homogeneous, the test of robustness of a statistical result across disparate samples is not the same as the typical statistical test of homogeneity. We must seek, instead, a prediction model that has general applicability. If a sample from Chicago appears to produce different results from a sample from Dallas, we cannot conclude that the study failed to replicate because varying environmental factors and genotype-by-environment interaction effects are likely responsible for the difference. But those environmental interactions may be too complicated or subtle to ever fully understand, and thus the lack of robustness across cities does have importance. Finally, large cohort studies and clinical trials cost tens of millions of dollars. It is not feasible to simply do a complete repeat of such massive studies. Potential users of the HapMap results for disease association studies would be wise to create networks of collaborating investigators to facilitate the evaluation and validation of predictive models. Just as the computer industry saw a sea change in the speed of progress when open source programming became widespread, the field of human genetics needs to lose its culture of competition and secrecy and replace it with protocols for sharing data, population samples, and analytical resources to accelerate the discovery of robust results.

Moving forward

Technical advances of the Hapmap project have been impressive, driving down the cost of SNP genotyping while driving up the quality and completeness of the data to a degree far exceeding initial expectations. It has provided us with an unprecedented genome-wide picture of human variation and linkage disequilibrium. But at one level, it is only a collection of data. Even the primary analyses to come from the project—on linkage disequilibrium, population recombination rates, finding recombination hotspots, and inference of natural selection—provide scant guidance in designing association tests that successfully accommodate the above challenges. We emphasize that in order to use the HapMap data optimally, we are forced to make crucial assumptions about the population-level behavior of polymorphisms and about the genetic etiology of disease risk. We must be open about these assumptions, and wherever possible we must seek to test them. Establishing whether the associations we seek are robust to these assumptions is central to the success of the application of the results of the HapMap effort.

There is no single optimal design that will most efficiently identify genetic and environmental factors that contribute to complex disease risk. Each study must wrestle with these challenges. Disease-oriented medical investigators who think that just by genotyping their cohorts they will produce a clean genomic solution to identifying genetic risk factors have been seriously misled. It seems inescapable that projects of larger scope will be better able to consider additional potential risk factors and identify relevant at risk subpopulations compared to the approach of simply genotyping the samples at hand. But a balance must be achieved, because huge studies that try to measure every factor are just as doomed as the overly narrow study designs that focus on the effects of only a few factors. In the end, it may be that the major impact we can have on public health is to show that alterations of the environment (e.g., diet, exercise, drug therapies) will provide the most cost-effective solutions to many of the major public health problems. Complex disease causation is almost certainly mediated by a network of effects, which means that neither genetic nor environmental agents are separate causes of the disease state, but instead it is their interactions that determine risk. It follows that the utility of genotypic information will be to identify those individuals for whom a change in environment would make the greatest difference in risk of disease.

Acknowledgments

This work was supported by NIH grants GM065509, HL072905, HL072810, and HL072904.

References

Anholt, R.R., Dilda, C.L., Chang, S., Fanara, J.J., Kulkarni, N.H., Ganguly, I., Rollmann, S.M., Kamdar, K.P., and Mackay, T.F. 2003. The genetic architecture of odor-guided behavior in *Drosophila*: Epistasis and the transcriptome. *Nat. Genet.* **35:** 180–184.

Brem, R.B. and Kruglyak, L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci.* **102:** 1572–1577.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305:** 869–872.

Frikke-Schmidt, R., Nordestgaard, B.G., Jensen, G.B., and Tybjaerg-Hansen, A. 2004. Genetic variation in ABC transporter A1 contributes to HDL cholesterol in the general population. *J. Clin. Invest.* **114:** 1343–1353.

Hamon, S.C., Stengård, J.H., Clark, A.G., Salomaa, V., Boerwinkle, E., and Sing, C.F. 2004. Evidence for non-additive influence of single nucleotide polymorphisms within the apolipoprotein E gene. *Ann. Hum. Genet.* **68:** 521–535.

Hinds, D.A., Seymour, A.B., Durham, L.K., Banerjee, P., Ballinger, D.G., Milos, P.M., Cox, D.R., Thompson, J.F., and Frazer, K.A. 2004. Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum. Genomics* **1:** 421–434.

Hirschhorn, J.N. and Altshuler, D. 2002. Once and again—Issues surrounding replication in genetic association studies. *J. Clin. Endocrinol. Metab.* **87:** 4438–4441.

Hirschhorn, J.N. and Daly, M.J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6:** 95–108.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., Sangiovanni, J.P., Mane, S.M., Mayne, S.T., et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308:** 385–389.

Peripato, A.C., De Brito, R.A., Matioli, S.R., Pletscher, L.S., Vaughn, T.T., and Cheverud, J.M. 2004. Epistasis affecting litter size in mice. *J. Evol. Biol.* **17:** 593–602.

Sawyer, S.L., Mukherjee, N., Pakstis, A.J., Feuk, L., Kidd, J.R., Brookes, A.J., and Kidd, K.K. 2005. Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.* **13:** 677–686.

Simpson, E.H. 1951. The interpretation of interaction in contingency tables. *J. Royal Stat. Soc. Series B* **13:** 238–241.

Sing, D.F., Stengard, J.H., and Kardia, S.L.R. 2004. Dynamic relationships between the genome and exposures to environments as causes of common human disease. *World Rev. Nutr. Diet.* **93:** 77–91.

Tu, Y.K., West, R., Ellison, G.T., and Gilthorpe, M.S. 2005. Why evidence for the fetal origins of adult disease might be a statistical artifact: The "reversal paradox" for the relation between birth weight and blood pressure in later life. *Am. J. Epidemiol.* **161:** 27–32.

Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. 2005. Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6:** 109–118.

Weinberg, C.R. 2005. Invited commentary: Barker meets Simpson. *Am. J. Epidemiol.* **161:** 33–35.

Weiss, K.M. and Terwilliger, J.D. 2000. How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26:** 151–157.

Zondervan, K.T. and Cardon, L.R. 2004. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5:** 89–100.

Zwick, M.E., Cutler, D.J., and Chakravarti, A. 2000. Patterns of genetic variation in Mendelian and complex traits. *Annu. Rev. Genomics Hum. Genet.* **1:** 387–407.