A *bias*-ed assessment of the use of SNPs in human complex traits

Joseph D Terwilliger*, Fatemeh Haghighi†, Tero S Hiekkalinna‡ and Harald HH Göring§

Although many biotechnological advancements have been made in the past decade, there has been very limited success in unraveling the genetic component of complex traits. Heavily invested research has been initiated based on etiological models of unrealistic simplicity and conducted under poor experimental designs, on data sets of insufficient size, leading to an overestimation of the effect sizes of genetic variants and the quantity and quality of linkage disequilibrium (LD). Arguments about whether families or unrelated individuals provide more power for gene mapping have been erroneously debated as issues of whether linkage or LD are more detectable sorts of correlation. Although the latter issue may be subject to debate, there is no doubt that family-based analysis is more powerful for detecting linkage and/or LD. If the recent advances in biotechnology are to be exploited effectively, vastly improved study designs will be imperative, as the reasons for the lack of success to date have much more to do with biology than technology, an issue that has become increasingly clear with the findings of the past years.

Addresses
*Columbia University, Department of Psychiatry and Columbia Genome Center, New York State Psychiatric Institute, Division of Molecular Genetics 1150 St Nicholas Avenue, Room 520-C New York, New York 10032, USA; e-mail: jdt3@columbia.edu
†Columbia Genome Center, 1150 St Nicholas Avenue, New York, New York 10032, USA; e-mail: fgh3@columbia.edu
‡National Public Health Institute, Department of Molecular Medicine, PO Box 104, FIN-00251, Helsinki, Finland;
e-mail: Tero.Hiekkalinna@helsinki.fi
§Department of Genetics, Southwest Foundation for Biomedical Research, PO Box 760549, San Antonio, Texas 78245-0549, USA;
e-mail: hgoring@darwin.sfbr.org

Current Opinion in Genetics & Development 2002, 12:726–734

0959-437X/02/$ – see front matter

DOI 10.1016/S0959-437X(02)00357-X

Abbreviations

CVCD	common variants for common diseases
HGP	human genome project
LD	linkage disequilibrium
OR	odds ratio
SNPs	single nucleotide polymorphisms

Introduction

The sequencing of the human genome has been a feat of industrial engineering that would have made Henry Ford, father of the assembly line, proud. However, it remains to be seen whether the promised scientific and public health advances are forthcoming. The proponents and leaders of the United States Human Genome Project (HGP) have described it as a series of 'five-year plans' to develop tools for application in unraveling the ephemeral relationships between genotypic and phenotypic variation in humans [1–4]. The first 'five-year plan' focused on the generation of genetic and physical maps, emphasizing the highly informative microsatellite markers that are most useful for gene mapping. The second plan focused on the sequencing of the genome, as a means of helping investigators to select 'positional candidate genes' within regions where they find evidence of linkage, and to develop the technologies needed for efficient re-sequencing, polymorphism screening, and genotyping within regions of interest. Initially these plans were presented as goals for the entire project, yet they have been largely completed ahead of schedule — although the available draft sequence of the human genome can hardly be considered completed. It is clear that substantial benefits were realized in making genome-wide linkage scans and subsequent positional cloning more efficient and cost-effective for studies of simple 'Mendelian' diseases using the traditional paradigms [5•]. In addition, the detailed characterization of the human genome will enable comparative genomics approaches in regions of synteny with other species on a genome-wide scale. Although these tools have helped us identify and characterize the genetic components of many rare human diseases, they have not led to the predicted advances in our understanding of the etiology of complex human diseases.

In the light of the rapid advances in engineering that accelerated the completion of the goals of the HGP, a new 'five-year plan' was proposed: to identify and characterize high-density maps of single nucleotide polymorphisms (SNPs) with high heterozygosities across continental populations [6,7]. This step was motivated by the hypothesis that common variants play an important role in the etiology of common diseases (CVCD) [8], and by the greater power predicted for genome-wide association studies under such simple models of disease etiology [9]. Note here that genome scanning for LD has indeed been a successful tool exploited in mapping genes that underlie Mendelian diseases, where such simple models are relevant, over the past decade in populations like Finland (see [10]). At the time, much debate ensued [11], with many investigators claiming that (unpublished) successes had already been achieved or were just around the corner [11]. However, to date, there are very few, if any, published success stories [12,13]. Now that hundreds of thousands of SNPs have been identified and characterized, the next plan aims, in part, to generate a so-called haplotype map (or 'HapMap') focused on identifying 'blocks' of linkage disequilibrium (LD) among tightly linked SNPs that are common across continental populations [14,15]. Such a HapMap, we are told, will allow us to carry out genome-wide association studies at greatly reduced cost, as markers could be chosen more wisely

in order to capture the haplotype structure of the genome, making genetic analysis in spreadsheets the potential future of human genetic epidemiology [15–18]. It remains to be seen, however, how many complex traits will be influenced by such common variants — with large marginal effects — that might be amenable to mapping with this tool, given that most genes mapped to date (in human as well as model organisms!) have been characterized by a wide range of common and rare variants each with minimal individual attributable risk [5•,19,20••]. Essentially, a major investment of capital is being made based on the hypothesized CVCD model, on the assumption that technology is the problem rather than biology [1,11–13]. If the problem is really biological complexity, as we hypothesize, then a reorientation of investment and research effort towards the development and implementation of study designs that may be more robust against violations of the current assumptions (particularly the assumption that the CVCD model is appropriate) may be advisable. The contributions of population geneticists to this discussion will be more and more important as time goes on, and it is hoped that they will continue to take an active interest in both the questions of population history and LD in human populations as well as in the evidence for different hypothetical etiological architectures for complex traits, not to mention in developing and implementing more powerful study designs.

The scientific value of a completed and polished human genome sequence is not in question — although the euphoria surrounding the announcement of the initial draft of the genome was a clear exaggeration. That said, the value of the most recent 'five-year plan' for developing the SNP and HapMaps is dubious at best, if the CVCD model does not turn out to be the rule. Much has been written in recent years to highlight the myriad reasons for adopting a more cautious perspective on the prognosis of gene mapping in complex traits and its likely effects on society. Readers who are not familiar with the details of this debate are encouraged to consult the following review papers on this topic written by ourselves [19,21••,22•,23•,24••] and others [12,13,20••,25•,26••,27••,28–35,36••,37•,38,39,40•,41,42,43••]. We do not intend to regurgitate our earlier arguments here, because nothing has happened since their publication to ameliorate those concerns, despite the continuing charge forward…

Technology versus biology

All gene-mapping studies aim to detect correlations between the genotypes of marker loci of known genomic locations and the phenotype of interest (which is often a disease). Genotypes of marker loci can be correlated with genotypes of the 'phenotypically active' locus under investigation because of linkage, which causes alleles of marker and phenotypically active loci to be co-transmitted in meiosis with high probability. Over many generations in a given population, the effects of tight linkage can be observed among distantly related individuals sharing some common ancestors (see [23•]). The consequences of such 'linkage in the population' are referred to as linkage disequilibrium (LD). It is essential to remember that linkage and LD are correlations among genotypes of loci, independent of phenotype. The advances in biotechnology will eventually lead to the correlations between some markers and any phenotypically relevant locus being arbitrarily tight, especially since we will soon be able to get complete sequence data for anyone we would wish to study, but no matter how tight these correlations are, the relationships between the genetic variation and phenotype remain ephemeral at best [30–33,35,36••]. The ultimate difficulties of gene mapping lie in the weakness of this latter correlation, and its properties. The reason for the difficulties in unraveling the genetics of complex disease is not a problem with the technology, but rather the complexities of the biology itself [5•,21••,32,37•].

In terms of probabilities, the problem in gene mapping is equivalent to testing the null hypothesis that $P(\mathbf{G_M},\mathbf{Ph}) = P(\mathbf{G_M})P(\mathbf{Ph})$, where $\mathbf{G_M}$ represents the vector of observed marker locus genotypes for all individuals in the study, and $\mathbf{Ph}$ represents the vector of observed phenotypes for all individuals in the study [21••,22•,23•,44–46,47••, 48,49••,50]. This representation highlights the conceptual issues and hides the computational complexity of evaluating these probabilities. If we let $\mathbf{G_P}$ represent the vector of genotypes at some locus which is influencing the trait, then

$$P(\mathbf{G_M},\mathbf{Ph}) = \sum_{G_P} P(\mathbf{G_M}|G_P)P(G_P|\mathbf{Ph}) \qquad (1)$$

where $P(\mathbf{G_M} \mid \mathbf{G_P})$ represents the correlations among genotypes of marker and trait loci due to linkage and/or LD, and $P(\mathbf{G_P} \mid \mathbf{Ph})$ is the detectance, or the predictive value of the observed phenotypes on the underlying genotypes under the ascertainment scheme employed in the study. These conditional probabilities are orthogonal to one another, meaning that their effects on the outcome of any study are independent. There must be a strong correlation between marker and trait locus genotypes (i.e. strong linkage and/or LD) and there must be a strong predictive relationship between the ascertained phenotypes and the underlying genotypes for a mapping study to be successful. The goal of the HGP is to provide scientists with technological tools to help them maximize the potential strength of $P(\mathbf{G_M} \mid \mathbf{G_P})$ in their studies. However, unless study designs are employed that maximize $P(\mathbf{G_P} \mid \mathbf{Ph})$, the failures in complex disease gene mapping will be perpetuated. Great technology cannot salvage a study if the detectance (in an ascertained sample) is weak and the study design is poor.

Note that the 'detectance' is implicitly conditional on the ascertainment scheme employed in a study, highlighting the importance of careful attention to study design. However, epidemiologists are traditionally most concerned with getting accurate estimates of the 'penetrance' or 'odds ratio' (OR), whereas geneticists are more concerned with detecting a genetic variant that is involved in disease etiology. Another way to view these two objectives is as

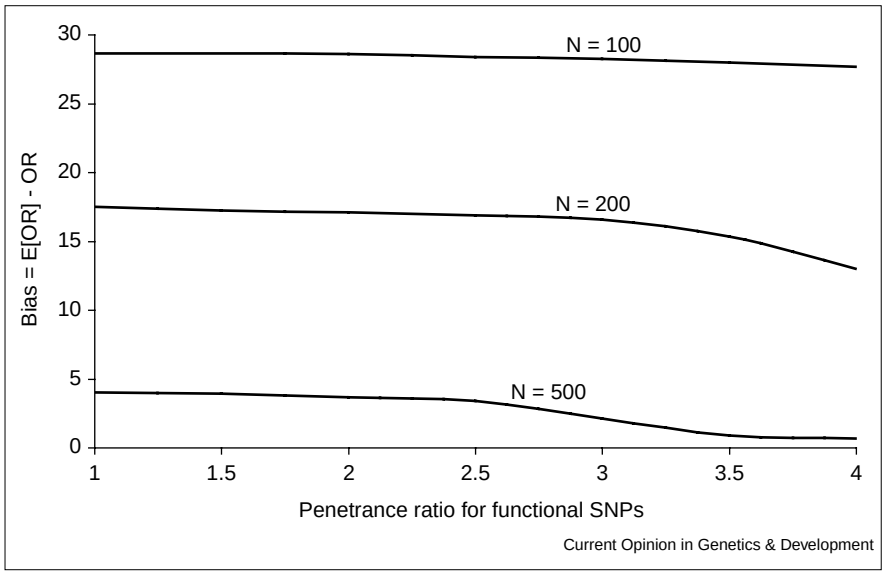
prediction and *inference*. Inference is the discovery of genetic risk factors, prediction the medically relevant goal. One element of prediction is related to confirmation of mutational effects in available samples, but in biomedicine the promised objective is accurate prediction of disease risk, often needing to be made many decades in advance. The optimal study designs for each of these goals are quite different [24••], because to estimate penetrance and ORs accurately, random ascertainment is optimal, whereas for detection of genetic variants which can influence a trait, ascertainment bias is imperative in order to enrich the sample for the risk factor being sought [24••]. This is an enormous difference of paramount importance in designing genetic epidemiology studies. Clearly random population cohorts will be indispensible for estimating the penetrance functions of genetic polymorphisms that are known, but they are not powerful for gene identification or mapping. Similarly, studies with enormous ascertainment bias are imperative for powerful gene mapping of variants with weak marginal effects, but are not useful for estimation of the effect size of the identified locus. No single study can accomplish both goals (i.e. detecting a risk gene and estimating its effect size), just as no epidemiological study can detect an environmental risk factor and accurately quantify the associated population risk of that factor unless the statistical significance is astronomical. This is a result of the effects of multiple testing, and of the fact that when power is low (as it clearly is in most existing complex disease-mapping studies), one must be lucky to detect an effect at all, and in such fortuitous experiments, the effect size of the identifiable factors must be inflated, compared to a random sample, as statistical significance is a direct function of the sample-specific effect size (see below). Only when significance far exceeds the multiple testing effects can accurate effect-size estimates be made (e.g. in a linkage study, this requires pointwise significance approaching $p < 10^{-10}$ rather than $p < 0.05$ [51••,52]).

If the genotypes of the functional variants are not available — which is still the case in almost all gene-mapping studies — then the second factor in the probability model from above comes into play, though it is important to highlight that this is of secondary importance to the power of a gene-mapping study. $P(\mathbf{G_M} \mid \mathbf{G_P})$ is the correlation between the (unseen) trait locus genotypes and the observed marker locus genotypes, and is a function of linkage and/or LD. Much discussion in past years has centered on the strengths and nature of these correlations among SNPs throughout the genome, largely as a result of the HGP's successes in sequencing the genome and in cataloguing the universally common variants across populations in small samples. We believe that the central issues that govern the success or failure of mapping studies are based on the detectance probabilities described above, and not on the correlations among loci as a result of linkage and/or LD. Admittedly, this used to be an important problem when studying diseases of high detectance with poor technology, but it is no longer the most significant impediment. In fact, we would be willing to posit, as we have in the past [22•,23•], that within what will seem to be a surprisingly short period of time, biotechnology will progress to the point where we will be able to generate complete sequence data for whatever genomic regions we are interested in looking at quickly and cheaply enough to make the whole LD issue essentially moot. Nonetheless, because the issue of SNP–SNP LD dominates much of the arguments of the proponents of the SNP map and the HapMap projects, these topics warrant some critical examination as well. We wish to emphasize, however, that as important as these issues may be, they are not central to why gene mapping has not been revolutionized as promised by the successful completion of the HGP's sequence of 'five-year plans'. Even if one has complete sequence data, the problems we consider to be of primary importance remain.

Study design: 'If it ain't tough to get, it ain't worth having' [53]

The utility of the existing SNP maps and of the haplotype maps which are to be generated in the next phase of the HGP is often said to lie in the hypothesis that they enable genome-wide LD mapping, typically on unrelated samples of singleton individuals with and without a particular disease; this has been perceived by many to be more powerful (or at least easier) than linkage studies on pedigrees ascertained to have multiple affected pedigree members. The issue, however, is incorrectly posed as a competition between LD and linkage-based methods. In fact it is a study design and ascertainment issue — families versus 'unrelated' (actually, as are all humans, very distantly related) individuals. It is clear that even if one wishes to do LD-based mapping with candidate genes, it is always more powerful to do this using pedigree data rather than unrelated individuals [23•,24••]. This is true because affected relatives are more likely to share some genetic risk factor than affected non-relatives, and the more affected individuals in a given pedigree, the greater the ascertainment bias in favor of genetic factors, as opposed to the environmental factors which we all acknowledge must be of paramount importance to traits which typically have heritabilities much less than 50%. Furthermore, genetic risk factors have a predictable correlation structure in exposures within families, much more so than environmental risk factors. This correlation in exposure provides an additional source of power in mapping studies, which becomes more and more valuable as pedigree size increases. Anyone who believes LD mapping will be more powerful than linkage mapping is essentially saying that the pedigree structures typically available for genetic epidemiology investigations are too small, and as a consequence they need to rely on ancestral meioses connecting individuals together in the population to make the pedigrees sufficiently large for powerful linkage mapping; yet it is clear that two affected relatives are much more likely, under any genetic model, to share the same risk variant than two affected "non-relatives". Statistical technologies and study designs for joint linkage and LD analysis on a wide range of data structures simultaneously have recently been described, and shown to be

Figure 1



10,000 replicates of a genome scan with 30,000 SNP markers were simulated, assuming a disease prevalence of 5%, risk allele frequency of 5%, and penetrance ratios k = P(Affected | DD or D+) / P(Affected | ++) ranging from 1 (no effect on the trait) to 4 (four-fold elevated risk for gene carriers). A case-control study was simulated, in which identical numbers of cases and controls were assumed, ranging from 100 of each to 500 of each, which is compatible with most studies in progress at the present time. The statistical significance of the case-control test was maximized over all markers in the genome, and the odds ratio was estimated at the most significant marker in the genome. For this marker, the bias was computed in each replicate, and was averaged over all replicates, with the resulting mean bias graphed in the figure. Note that even for sample sizes as large as 500, the bias does not begin to decline until the penetrance ratio is at least 2, which is higher than the expected penetrance ratios for most loci likely to be implicated in the aetiology of complex traits.

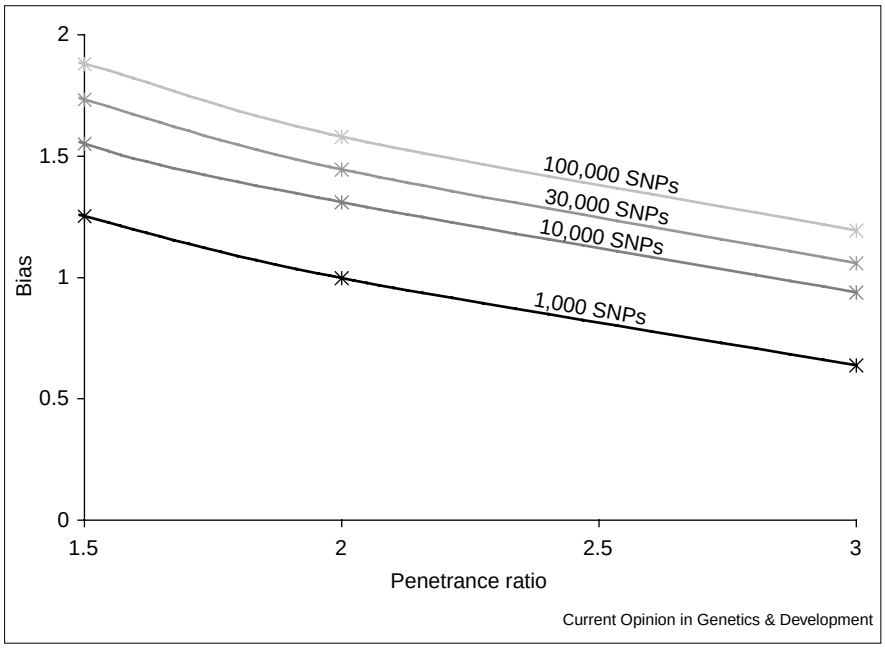
uniformly more powerful than pure linkage-based or association-based approaches [47••], bringing into play the common-sense statement that genetics is more powerful when family data are used because the power of genetics is based entirely on the correlation structure in risk factor exposure among related persons [*22*•]!

In particular, user-friendly computer software for performing such joint linkage and LD analyses on a variety of data structures with qualitative traits, 'Pseudomarker', has been developed recently [54], and has contributed to some discoveries in recent months. One example where this method has been efficiently used was in the discovery of a single common variant that seems to be related to adult lactase persistence across a wide variety of European populations [55••]. This was, however, a fully penetrant trait for which everyone in the study population with the trait carried clonal copies of a single ancestral variant. Although this might suggest grounds for some optimism and does provide an example of a single common variant influencing a common trait, it should be pointed out that this variant was located in an intron of a nearby gene, unrelated to the gene whose regulation the variant controls. Indeed this variant is located so far from the gene (~15 kb) that most candidate gene investigations would not have even looked there. The only reason it was detectable is because the phenotype predicted the genotype at this variant with complete determinism, which is not the expected model for the complex multifactorial diseases that are the target of most LD studies. Furthermore, it should be pointed out that in this instance, as in most instances on the record in which LD has been used successfully, strong linkage and LD were detected using microsatellite loci, without the need for SNPs, consistent with moves in many quarters away from SNP-based LD mapping back towards microsatellite-based mapping approaches [56,57•,58,59]. In passing, it should be pointed out that there are numerous successful applications of genome-wide LD mapping, virtually all of which were done for rare Mendelian disorders in population isolates in which it was true that all affected individuals shared disease alleles which were identical-by-descent from a common ancestor (e.g. [10]), whereas similar efforts for complex diseases have not been successful (e.g. see [60]).

Biased effect size estimates

As already alluded to above, no matter what ascertainment scheme is used, nor whether one is studying linkage or LD, it is not possible both to localize a gene and to accurately estimate its effect size on a single dataset. The reason is that even if the effect size could be estimated from a given sample without bias in a single pointwise analysis, it will be biased upwards, when the estimation is made conditional on there being significant evidence of correlation, or conditional on the marker being the most significant out of a large number of tests performed in the search. This is because the test statistic itself is a direct function of the parameters describing the effect size [51••,52]. We have previously demonstrated for genome-wide linkage studies how severely the effect size of a locus is overestimated conditional on a significant linkage finding being obtained, even if the sample is randomly ascertained and thus fairly represents the study population as a whole [51••,61]. For genome-wide LD studies with discrete traits, this bias is even more severe, as a result of the increased multiple testing problem and concomitant lowering of the

Figure 2



1,000,000 replicates of a case control study with 100 cases and 100 controls were simulated for a single functional variant with population frequency of 30% – consistent with the CVCD hypothesis – for a trait with 5% prevalence, and penetrance ratios ranging from 1.5 to 3. In this case, the odds ratio was computed conditional on a given replicate leading to statistically significant evidence of association, after correction for multiple testing. The number of markers assumed to be tested ranges from 1000 to 100,000 in the figure, with the average bias in the pointwise effect size estimate increasing with the number of markers typed in the genome scan, and decreasing only slowly as the penetrance ratio increases, in an almost linear manner.

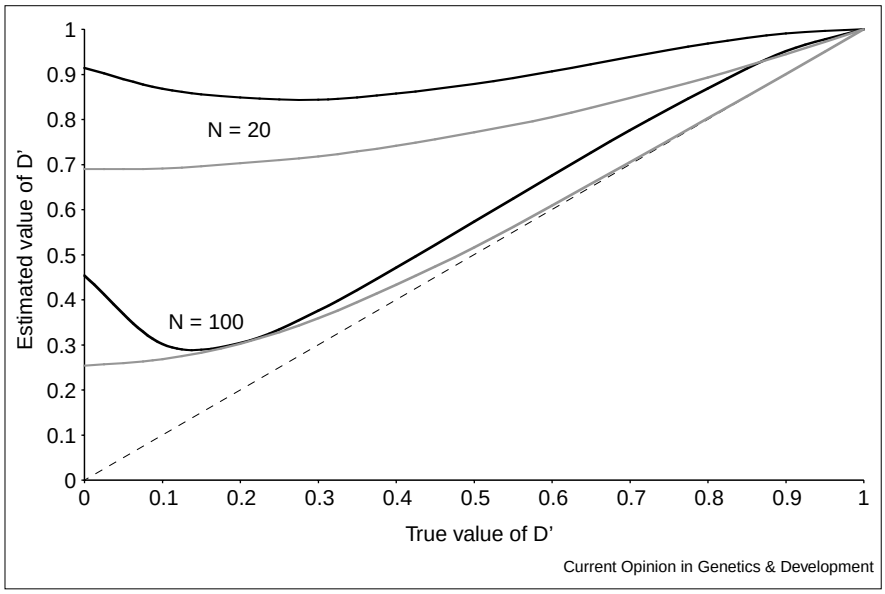
required point-wise significance. Of course, the same problem plagues all epidemiology studies where multiple hypotheses involving multiple candidate risk factors are considered in one study, although this is typically ignored in practice, leading to a plethora of false positive claims and irreproducible results in the epidemiological literature as well.

Figure 1 shows the effect of maximizing the evidence for allelic association over 30,000 markers in a genome scan on the estimated odds ratio for varying sample sizes. In this case, both prevalence and disease allele frequency were assumed to be 0.05, and a dominant mode of inheritance was assumed, with a variety of penetrance ratios k = P(Affected | DD or D+) / P(Affected | ++). In the case where the sample size is either 100 or 200 cases and controls, the estimated odds ratio is virtually independent of the true mode of inheritance, and is grossly overestimated to be many times its true value. When the sample size is as high as 500 cases and controls, there is still a rather significant bias, where a true OR of two is estimated to be closer to five. In Figure 2, the effect of the number of markers typed in a genome scan is indicated, where here the bias is shown in the OR at the functional SNP itself, conditional on there being significant evidence of association, regardless of whether another genomic location was more significant than the true SNP. In this case, a much more common variant, with disease allele frequency of 0.3, is assumed, and disease prevalence remains at 0.05. In fact, the mean OR estimates conditional on test significance are much larger than the true value in the population, and the mean OR estimates for true positives (i.e. at a functional SNP) and false positives (i.e. at a non-functional SNP) are quite similar to each other. The effect-size estimate can thus be essentially independent of the true state of nature! Corollaries of this observation are that, first, one cannot distinguish true and false positives from each other on the basis of their effect size estimate and, second, in studies of low power one can predict from the outset what the estimated effect-size estimate will be, if a significant test were achieved somewhere in the genome. Contrary to the arguments of Allison *et al.* [52], we continue to argue that efforts in bias correction are unsatisfactory, and propose estimation of the locus effect in an independent dataset as the only suitable means of obtaining satisfactory estimates.

In contrast to many epidemiologists, geneticists often care much more about the power to map genes, and less about the accuracy of the obtained gene effects. However, the failure to appreciate the large upward bias in effect size estimates at peak mapping statistics in the genome often unwittingly results in serious frustration to geneticists who try in vain to replicate an earlier finding of linkage and/or LD. The reason is that the upwardly biased estimates of genetic effect published in a study are often taken, inappropriately, as input values for sample size computations for subsequent replication studies. As a result, the sample size requirements are often vastly underestimated, thus leading to frequent failure of replication even in those instances where the initial finding pointed to a true functional gene [51••,61]. It is sometimes said that the sample size for replication needs to be much larger than for the initial finding. This is, of course, not formally correct, as the power of a study does not depend on whether it is the first of its kind or not. Rather, there is probably a significant publication bias in the sense that studies with a statistically significant finding are more likely to be published. The

Figure 3

A simulation study was conducted to determine the degree of small sample bias in the estimation of the LD measure D′ (see [65]). For purposes of this simulation study, sample sizes of 20 chromosomes (as was used in [62]) and 100 chromosomes (slightly larger than that used in [14]) were used. Two loci were compared, the first with allele frequency of 0.1, and the second with allele frequencies of either 0.1 (dark lines in figure) or 0.5 (Grey lines in figure), over the entire positive range of D′. The average of the estimated D′ values over all replicates is graphed on the *y*-axis with the true D′ indicated on the *x*-axis. There is a dramatic upward bias throughout the range of D′ other than when D′ = 1, at which point the estimate must be unbiased by definition.



initial report of a locus is thus likely based on a 'lucky' study. Replication on a separate dataset of equal size to the original study will thus most likely fail, and many such replication attempts will be required until another one turns out to be equally 'lucky', unless the sample size is increased dramatically. By this time, even very weak second hits are considered to confirm that a gene is causal, and there seems to be a sense that this also means that it is an *important* causal factor, rather than one with much less impact than originally estimated. This is despite the fact that for some diseases, like schizophrenia, there are reported positive linkage findings on virtually every arm of every chromosome [22•], making the results of any new study a 'replication' of at least one of the previously reported findings in some sense — an additional sort of "multiple-testing" problem which needs to be dealt with.

Bias in linkage disequilibrium estimation

Not only are the estimated effect-sizes of newly mapped genes grossly overestimated, but also the very level of LD itself, which is the basis for much of the euphoria surrounding the HapMap project. Numerous investigators have been looking at the strength of LD in a variety of genomic regions based on very small samples of anywhere from 20 to 100 chromosomes [14,62]. Such sample sizes are so small that it is well known that estimates of LD will be systematically biased upward, often to a great degree [63•,64]. Although bias-correction methods have been proposed and applied in practice ([64]; T Varilo *et al*., unpublished data) the biases are often ignored. We believe that by the time the HapMap is finished, biotechnology advances, with concomitant reduction in cost of large-scale resequencing, will have made the degree of LD as well as the HapMap itself largely irrelevant to gene mapping, but it is important to point out that many of the estimates of LD are inflated (see also [37•]).

Figure 3 shows the amount of bias in estimates of D′ (a measure of the strength of LD between two SNPs, with 0 being no LD and 1 being complete LD — see [65] for details) from samples of 20 and 100 chromosomes, which approximate the sample sizes used by [62] and [14], to show the enormity of the problem. Furthermore, the definition of 'blocks' used by these authors is likewise somewhat susceptible to misinterpretation. Note that under the definition of a block used in [62] almost all pairs of SNPs will constitute a 'haplotype block' even in the absence of LD, and that 52% of all haplotype blocks identified in that paper comprise only one or two SNPs per block. The definition of 'strong LD' used in [14] was that $|D'| > 0.5$, meaning that one observes about 50% of the maximum amount of LD that a given pair of loci could theoretically exhibit. However, the probability of getting an estimate of $|D'| > 0.5$ from a sample of 100 chromosomes can be quite large, even when there is virtually no LD — for example, in simulation studies (data not shown) for two SNPs, each with minor allele frequency of 0.1, roughly 40% of the time you will have a naïve point estimate of $|D'| > 0.5$, in the absence of bias correction, because of the same small-sample bias described above.

Conclusions

In conclusion it may seem odd, in a review paper about advances towards an understanding of the etiology of complex traits through the advent of the HapMap and other technological advances, that we dwell so much on the underlying rationale behind these projects, especially given that it is almost a foregone conclusion that they will proceed. We decided to emphasize these issues largely because there have been virtually no discoveries in this area which seem likely to improve our understanding of complex disease genetics, unless the CVCD hypothesis

turns out to be the rule, rather than the occasional exception. Much of the literature in this area over the previous year consists of speculation and opinion rather than new empirical evidence or theoretical models. In fact, despite most of the theoretical and empirical work in the previous years being consistent with our more skeptical outlook, these projects seem to be racing ahead with little focused debate. The danger here is the widespread propagation of the CVCD hypothesis and large-scale case-control studies as the "accepted" approach for mapping complex disease genes, despite the substantial empirical and theoretical evidence against them. Without careful consideration of the underlying models, it is easy to be swayed in this direction, especially with the ready availability of funding at the moment in these areas of research. However, ten years down the road, small labs may be in trouble if they get too caught up in copying the strategies of the large genomics factories yet no public health benefits arise from the proposed case-control approach to genetic epidemiology. For a large factory to take this approach may be logical, as they will be seen as successful if one gene can be mapped out of 100 studies they undertake, but a lab with resources focused on one trait of interest has to rely on good fortune, and hope their trait has a single common variant responsible for a large amount of the population risk of disease, which is a risky gamble at best [19,20••,21••,32,37•,66].

It is natural to be somewhat concerned whenever government officials and their sycophants start talking in terms of sequential five-year plans [2] that promise to "save lives the world over" [1]. On paper, the statistics of industrial output seem to exceed prognosticated achievements, whereas in reality the purported goal — in this case, of understanding the relationships between genetic variation and common disease — seems no closer despite these technological advances [2]. For this reason, some re-alignment of the grand plan to the true aims at regular time intervals should not seem imprudent. And when such enormous resources become collectivized in a small number of large industrial complexes, is it really to the benefit of science, let alone public health? Historically, many of the biggest scientific advances have come when there was less funding and more thinking, the relative proportions of which have changed dramatically in recent years. We suggest that before getting too deep into the promises of the HGP, it may be wise to step back momentarily and make sure that we are not falling into a convenient trap and charging imprudently into expensive blind alleys.

Update

A new paper [67••] investigates the likely allelic architecture of a locus involved in susceptibility to some common diseases. The authors demonstrate from population genetics principles why the CVCD hypothesis is unlikely to hold in general, looking at best-case-scenario models under which one might hope that the CVCD model could hold. The authors' purely gene-based argument shows that even in the most propitious circumstances the CVCD model is unlikely to be relevant to most practical situations, despite the conciliatory tone the authors take in the discussion section, which ignores the confounding issues of further etiological and biological complexity.

Acknowledgements

Grants MH63749, MH59490, HL45522, HL28972, GM31575, MH59490 from the National Institutes Health and Alfred E. Sloan/DOE training grant DE-FG02-00ERG2970 are gratefully acknowledged, as is support from the Emil Aaltonen Foundation. Thanks to Joseph H Lee and Kenneth M Weiss and the editors of this issue for critical reading and comments on the manuscript.

Software note

The pseudomarker program for joint linkage and LD analysis on a variety of data structures, including case-control data, nuclear family data, and extended pedigree data are available from the authors at http://www.helsinki.fi/~tsjuntun/pseudomarker/index.html

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Couzin J: New mapping project splits the community. *Science* 2002, 296:1391-1392.

2. Collins F, Galas D: A new 5-year plan for the United-States Human Genome Project. *Science* 1993, 262:43-46.

3. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L, Fearon E, Hartwelt L, Langley CH, Mathies RA *et al.*: New goals for the US Human Genome Project: 1998-2003. *Science* 1998, 282:682-689.

4. Collins FS, McKusick VA: Implications of the Human Genome Project for medical science. *J Am Med Assoc* 2001, 285:540-544.

5. • Weiss K: Goings on in Mendel's Garden. *Evol Anthropol* 2002, 11:40-44.

This commentary discusses the misappropriation of Mendel's concepts by those studying traits with complex modes of inheritance. As the author points out, Mendel himself recognized the difference, and realized that the only traits which could be understood in a straightforward manner were those that were deterministically influenced by specific genes.

6. Collins FS, Guyer MS, Chakravarti A: Variations on a theme: cataloging human DNA sequence variation. *Science* 1997, 278:1580-1581.

7. Collins FS, Mansoura MK: The human genome project: revealing the shared inheritance of all humankind. *Cancer* 2001, 91:221-225.

8. Reich DE, Lander ES: On the allelic spectrum of human disease. *Trends Genet* 2001, 17:502-510.

9. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996, 273:1516-1517.

10. Nikali K, Suomalainen A, Terwilliger J, Koskinen T, Weissenbach J, Peltonen L: Random search for shared chromosomal regions in 4 affected individuals – the assignment of a new hereditary ataxia locus. *Am J Hum Genet* 1995, 56:1088-1095.

11. Pennisi E: A closer look at SNPs suggests difficulties. *Science* 1998, 281:1787-1789.

12. Patterson M: That damned elusive polygene. *Nat Rev Genet* 2000, 1:86-86.

13. Patterson M: Wake-up call for genome scanners. *Nat Rev Genet* 2002, 3:9.

14. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R *et al.*: Linkage disequilibrium in the human genome. *Nature* 2001, 411:199-204.

15. Olivier M, Bustos VI, Levy MR, Smick GA, Moreno I, Bushard JM, Almendras AA, Sheppard K, Zierten DL, Aggarwal A *et al.*: Complex high-resolution linkage disequilibrium and haplotype patterns of

single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21. *Genomics* 2001, 78:64-72.

16. Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001, 29:229-232.

17. Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F *et al.*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001, 29:233-237.

18. Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC: How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* 2002, 3:379-391.

19. Terwilliger JD, Weiss KM: Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* 1998, 9:578-594.

20. •• Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001, 69:124-137.
One of the first legitimate attempts to resolve the debate about whether rare or common variants are more likely to be involved in human complex diseases from a sophisticated population genetics perspective. Of course, as one should expect, the author concludes by suggesting the truth will be a mixture of both forms of risk factors, the balance of which depends on difficult to measure parameters of population history. Nevertheless, it does make one question the validity of the CVCD model which underlies the proposed HapMap project.

21. •• Weiss KM, Terwilliger JD: How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000, 26:151-157.
This commentary provides an overview of the current issues influencing complex trait gene mapping, highlighting some of the problems and popular misconceptions which may lead to overestimation of the prognosis for unraveling the aetiology of multifactorial traits using naïve approaches.

22. • Terwilliger JD, Goring HHH: Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000, 72:63-132.
This review paper provides a general introduction to the conceptual underpinnings of gene mapping aimed at an audience of anthropologists. While heavy on concepts and light on mathematical detail, it surveys the field and reviews some of the existing data with insight into to why simple naïve study designs are likely to be doomed to failure in complex trait genetic studies.

23. • Terwilliger JD: On the resolution and feasiblity of genome scanning approaches. *Adv Genet* 2001, 42:351-391.
This paper attempts to quantify the resolution of linkage and LD mapping in the presence of complete identity-by-descent information under some simple mathematical models of population structure and history. Further, the effects of ascertaining multiplex families as compared to singleton-affected individuals on association studies is quantified and explained.

24. •• Terwilliger JD, Göring HHH, Magnusson PKE, Lee JH: Study design for genetic epidemiology and gene mapping: the Korean diaspora project. *Shengming Kexue Yanjiu (Life Science Research)* 2002, 6:95-115
An overview of the various study designs used in gene mapping and genetic epidemiology is presented, highlighting the advantages and weaknesses of each data structure and study design commonly used for estimating and detecting the effects of genetic, environmental, and cultural factors on the etiology of a given trait, with special emphasis on ways to combine data structures obtained with different ascertainment schemes in order to increase the accuracy of such prognostications.

25. • Millikan R: The changing face of epidemiology in the genomics era. *Epidemiology* 2002, 13:472-480.
This paper provides a nice synthesis of the epidemiologists' perspective on the genomics 'revolution'. Obviously, geneticists have a lot to learn from epidemiologists and vice versa and cross talk between both disciplines is imperative. We should try to study genetic and environmental risk factors jointly in common studies, using large pedigrees of related individuals rather than random population cohorts.

26. •• Holtzman NA: Putting the search for genes in perspective. *Int J Health Serv* 2001, 31:445-461.
This commentary discusses some of the false promises of the genomics era and describes the author's more realistic outlook on the situation, strongly questioning the enormous expenditure of the US government on the blind search for genomic approaches to public health problems

27. •• Mackay TFC: The genetic architecture of quantitative traits. *Annu Rev Genet* 2001, 35:303-339.
This is one of few attempts in the literature to glean information about the likely architecture of human complex traits by examining the question in a natural animal population. It makes one wonder why no-one is looking in earnest at subway rats to see if they can map any of the known rat quantitative trait loci in natural populations before making the huge jump from inbred rats to outbred 'subway' humans.

28. Peltonen L, Palotie A, Lange K: Use of population isolates for mapping complex traits. *Nat Rev Genet* 2000, 1:182-190.

29. Rees J: Complex disease and the new clinical sciences. *Science* 2002, 296:698-701.

30. Schlichting CD, Pigliucci M: *Phenotypic Evolution: a Reaction Norm Perspective.* Sunderland, MA: Sinauer Associates, Inc.; 1998.

31. Weiss KM: *Genetic Variation and Human Disease: Principles and Evolutionary Approaches.* Cambridge: Cambridge University Press; 1993.

32. Weiss KM: Is there a paradigm shift in genetics? Lessons from the study of human diseases. *Mol Phylogenet Evol* 1996, 5:259-265.

33. Weiss KM: In search of human variation. *Genome Res* 1998, 8:691-697.

34. Weiss KM, Clark AG, Fullerton SM, Taylor SL, Nickerson DA, Sing CF: Evaluating the phenotypic effects of SNP variation: sampling issues. *Am J Hum Genet* 1999, 65(Suppl S):8.

35. Weiss KM, Buchanan AV: Rediscovering Darwin after a Darwinian century. *Evol Anthropol* 2000, 9:187-200.

36. •• Weiss KM, Fullerton SM: Phenogenetic drift and the evolution of genotype-phenotype relationships. *Theor Popul Biol* 2000, 57:187-195.
A nice overview of the evolutionary forces which have moulded the genetic structure of contemporary populations. Clearly it is the phenotype which is constrained in evolution, rather than the genetic mechanism through which the phenotype is obtained. As the authors demonstrate clearly, phenotypic similarity is not incompatible with great genotypic dissimilarity, something that it is very important to consider when designing genetic epidemiology studies, which are likely to be characterized by great heterogeneity.

37. • Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002, 18:19-24.
In this review paper, more challenges to the CVCD model are presented, along with a discussion of the difficulties in accurately estimating the quantity and quality of LD from small samples. An overview of various measures of LD, and what they mean is also given. Anyone seriously interested in exploiting the power of LD for mapping should make themselves familiar with the population genetics concepts introduced in this review.

38. Wright AF, Carothers AD, Pirastu M: Population choice in mapping genes for complex diseases. *Nat Genet* 1999, 23:397-404.

39. Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001, 27:234-236.

40. • Brookes AJ: Rethinking genetic strategies to study complex diseases. *Trends Mol Med* 2001, 7:512-516.
Even the biotechnology folks are starting to question the power of SNP-based mapping approaches, after years of trying to work out some success stories. Nevertheless, this review presents a notably more optimistic viewpoint on the subject than our own.

41. Cardon LR, Bell JI: Association study designs for complex diseases. *Nat Rev Genet* 2001, 2:91-99.

42. Ghosh S, Collins FS: The geneticist's approach to complex disease. *Annu Rev Med* 1996, 47:333-353.

43. •• Horrobin DE: Realism in drug discovery — could Cassandra be right? *Nat Biotechnol* 2001, 19:1099-1100.
Pessimism is not confined to gene-mapping prospects; those involved in pharmacogenetics may have even greater difficulties, as this author points out. Most drug response traits are not even known to be heritable or genetic, and yet we have this blind faith that the genomics 'revolution' will give us technological solutions to these illusive problems. In some cases, maybe, but it is dangerous to assume this blindly, as the author clearly explains.

44. Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors I: Complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 2000, 66:1095-1106.

45. Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors II: Marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet* 2000, 66:1107-1118.

46. Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors III: Marker loci and their map as nuisance parameters. *Am J Hum Genet* 2000, 66:1298-1309.

47. Goring HHH, Terwilliger JD: Linkage analysis in the presence of
•• errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 2000, 66:1310-1327.
This paper shows how various 'model-free' test statistics used in linkage and LD analysis, such as affected sib-pair tests, haplotype relative risk tests, TDT tests, and case-control tests, are all equivalent to the same 'model-based' parametric likelihood ratio test applied to slightly different data structures. By combining data from all available data structures – from case-control data to large extended pedigrees – in a single analysis, a much more powerful and appropriate series of test statistics is proposed, for which software is provided by the authors.

48. Goring HHH, Ott J, Terwilliger JD: A common framework for model-based or model-free, twopoint or multipoint, linkage and/or linkage disequilibrium analysis of complex traits. *Am J Hum Genet* 1999, 65(Suppl S):1397.

49. Terwilliger JD: A likelihood-based extended admixture model of
•• oligogenic inheritance in 'model-based' and 'model-free' analysis. *Eur J Hum Genet* 2000, 8:399-406.
This paper debunks the mythology that one can gain power in linkage analysis from modeling epistatic interactions among multiple disease-predisposing loci, at least in the case of complex qualitative traits in human genetics.

50. Terwilliger JD, Zollner S, Laan M, Paabo S: Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion. *Hum Hered* 1998, 48:138-154.

51. Göring HHH, Terwilliger JD, Blangero J: Large upward bias in
•• estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001, 69:1357-1369.
In this paper, the bias in estimation of the locus-specific heritability in a genome scan, resulting from multiple testing, is described for variance-components linkage analysis. Clearly the same phenomenon described in this paper affects all methods of quantitative and qualitative linkage or LD analysis, as well as any epidemiological or other study in which multiple hypotheses are being tested.

52. Allison DB, Fernandez JR, Heo M, Zhu SK, Etzel C, Beasley TM, Amos CI: Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am J Hum Genet* 2002, 70:575-585.

53. Hatfield FC: *Power: A Scientific Approach.* Chicago: Contemporary Books, Inc.; 1989.

54. Hiekkalinna T, Göring HHH, Peltonen L, Terwilliger JD: PSEUDOMARKER: computer software for linkage and linkage disequilibrium analysis of complex traits. *Am J Hum Genet* 2000, 67:1836.

55. Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I:
•• Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 2002, 30:233-237.
Lactose-intolerance appears to be the normal state for humans, as all lactose-tolerant individuals studied across a variety of populations seem to share a common variant at one genomic location, which, frighteningly enough, is located in an intron of a gene 15kb away from the gene whose regulation it seems to control. While this is an example of a common variant for a common trait, which was mapped by joint analysis of linkage and LD in large extended pedigrees using microsatellite markers, it should send shudders down the spine of people involved in candidate gene studies, with respect to how far away from a gene a strong regulatory variant can be located!

56. Tamiya G, Oka A, Okamoto K, Endo T, Makino S, Hayashi H, Iizuka M, Tokubo E, Sato R, Takaki A *et al.*: A genome-wide association study of psoriasis vulgaris using polymorphic microsatellite markers and microarray technology. *Am J Hum Genet* 2001, 69:2232.

57. Okamoto K, Makino S, Endo T, Hayashi H, Oka A, Fujimoto K,
• Denda A, Watanabe H, Tokubo E, Sato R *et al.*: Comprehensive setting of 30,000 polymorphic microsatellite markers throughout human genome. *Am J Hum Genet* 2001, 69:1617.
In Japan, some investigators are proposing to perform genome scans with a dense map of 30,000 microsatellites rather than the 30,000 SNPs being proposed by the proponents of the HapMap. This is of special interest because the microsatellite map presents a potentially more powerful resource competing with the HapMap, as these markers may be as informative about LD as the SNPs in the HapMap, at least in many genomic regions, and at the same time, as they have a mix of rare and common alleles, they may be potentially useful in detecting rare variants as well as common ones in some cases.

58. Ohashi J, Tokunaga K: A comparison of cost-effectiveness between microsatellite and single nucleotide polymorphism markers in genomewide linkage disequilibrium testing. *Am J Hum Genet* 2001, 69:1345.

59. Ohashi J, Tokunaga K: The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J Hum Genet* 2001, 46:478-482.

60. Hovatta I, Varilo T, Suvisaari J, Terwilliger JD, Ollikainen V, Arajarvi R, Juvonen H, Kokko-Sahin ML, Vaisanen L, Mannila H *et al.*: A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet* 1999, 65:1114-1124.

61. Goring HHH, Terwilliger JD, Blangero J: Genome scans for quantitative trait loci using variance components linkage analysis: upward bias in heritability estimates attributable to individual quantitative trait loci at lod score peaks. *Am J Hum Genet* 2000, 67:1176.

62. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP *et al.*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001, 294:1719-1723.

63. Zapata C, Carollo C, Rodriguez S: Sampling variance and
• distribution of the D′ measure of overall gametic disequilibrium between multiallelic loci. *Ann Hum Genet* 2001, 65:395-406.
A theoretical description of the bias in estimation of D′ from small samples is provided by the authors. Those investigators basing their future study designs on such things as the amount of LD seen in comparisons of random SNPs should familiarize themselves with the details of the measures and associated biases to avoid making unadvised extrapolations.

64. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al.*: The structure of haplotype blocks in the human genome. *Science* 2002, 296:2225-2229.

65. Lewontin RC: On measures of gametic disequilibrium. *Genetics* 1988, 120:849-852.

66. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E *et al.*: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998, 63:595-612.

67. Pritchard JK, Cox NJ. The allelic architecture of human disease
•• genes: common disease-common variant... or not? *Hum Mol Genet* 2002 11:2417-2423.
See 'Update'.