How many diseases does it take to map a gene with SNPs?

Kenneth M. Weiss[1] & Joseph D. Terwilliger[2]

"They all talked at once, their voices insistent and contradictory and impatient, making of unreality a possibility, then a probability, then an incontrovertible fact, as people will when their desires become words."
—W. Faulkner, *The Sound and the Fury*, 1929

Through rose-coloured glasses darkly

There are more than a few parallels between the California gold rush and today's frenetic drive towards linkage disequilibrium (LD) mapping based on single-nucleotide polymorphisms (SNPs). This is fuelled by a faith that the genetic determinants of complex traits are tractable, and that knowledge of genetic variation will materially improve the diagnosis, treatment or prevention of a substantial fraction of cases of the diseases that constitute the major public health burden of industrialized nations[1–5]. Much of the enthusiasm is based on the hope that the marginal effects of common allelic variants account for a substantial proportion of the population risk for such diseases in a usefully predictive way[6,7]. A main area of effort has been to develop better molecular and statistical technologies[8–12], often evaluated by the question: how many SNPs (or other markers) do we need to map genes for complex diseases? We think the question is inappropriately posed, as the problem may be one primarily of biology rather than technology. Here we try to clarify fundamental issues related to LD-based mapping, for which high hopes are widely held. These issues have ramifications that may not be widely appreciated[13,14]. A balanced discourse on the prospects for LD mapping of complex traits might be gained by pointing out some of the potential problems and inverting the question to: how many diseases must we sort through to find individual alleles of widespread impact on disease? Are the current strategies the best we can do for public health, or even for genetics?

A SNP is not the same as a disease-predisposing allele

A linkage or LD analysis would test the null hypothesis that alleles of some gene (G_p) that influence some phenotype (Ph) are inherited independently of alleles at some specific chromosomal position (G_X; Box 1, equation 1). In this case, the only correlation to be tested is that owing to linkage or LD correlating G_X and G_P (Fig. 1, green arrow). As most SNPs have arisen only once in history (or only a few times, but on different chromosomal backgrounds), one could potentially map the position, X, of such a gene based on LD with markers at nearby chromosomal positions.

For genes influencing disease outcomes, until the locus is identified (which is the mapping objective), one only observes phenotypes (Ph), and not the underlying risk genotypes (G_P), such that inference can only be based on a test of independence of observed phenotypes and marker genotypes (Fig. 1, heavy black arrow, and Box 1, equation 2). As correlations between G_X and Ph only exist because of correlations between G_X and G_P (linkage and/or LD) *and* between G_P and Ph, the power of a study is a function of $P(G_P|Ph)$, the ability of the observed disease phenotypes to predict the underlying risk genotypes (Fig. 1, dotted purple arrow), conditional on the ascertainment. Because there is a one-to-many relationship between phenotypes and genotypes, one must allow for the possibility of all admissible genotypes of the trait locus, conditional on the observed phenotypes. That is to say we test whether or not the

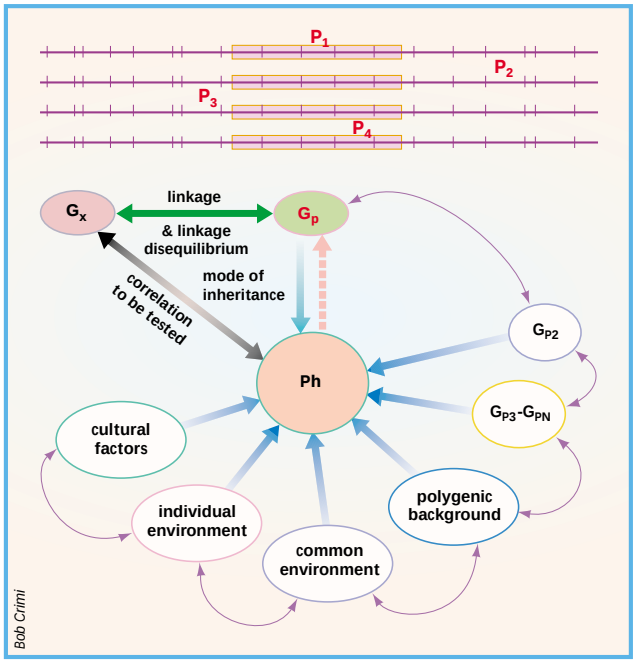

Fig. 1 Schematic model of trait aetiology. The phenotype under study, **Ph**, is influenced by diverse genetic, environmental and cultural factors (with interactions indicated in simplified form). Genetic factors may include many loci of small or large effect, G_{Pi}, and polygenic background. Marker genotypes, **Gx**, are near to (and hopefully correlated with) genetic factor, $\mathbf{G_p}$, that affects the phenotype. Genetic epidemiology tries to correlate $\mathbf{G_x}$ with **Ph** to localize $\mathbf{G_p}$. Above the diagram, the horizontal lines represent different copies of a chromosome; vertical hash marks show marker loci in and around the gene, $\mathbf{G_p}$, affecting the trait. The red $\mathbf{P_i}$ are the chromosomal locations of aetiologically relevant variants, relative to **Ph**.

[1]*Departments of Anthropology and Biology, Penn State University, University Park, Pennsylvania, USA.* [2]*Department of Psychiatry and Columbia Genome Center, Columbia University, New York, New York, USA. Correspondence should be addressed to K.M.W. (e-mail: kmw4@psu.edu).*

Box 1

equation 1

$$P(\mathbf{G_x},\mathbf{G_p})=P(\mathbf{G_x}|\mathbf{G_p})P(\mathbf{G_p})\stackrel{?}{=}P(\mathbf{G_x})P(\mathbf{G_p})$$

equation 2

$$P(\mathbf{G_x},\mathbf{Ph})=P(\mathbf{G_x}|\mathbf{Ph})P(\mathbf{Ph})\stackrel{?}{=}P(\mathbf{G_x})P(\mathbf{Ph})$$

equation 3

$$P(\mathbf{G_x},\mathbf{Ph})=\sum_{G_p}P(\mathbf{G_x}|\mathbf{G_p})P(\mathbf{G_p}|\mathbf{Ph})P(\mathbf{Ph})\stackrel{?}{=}P(\mathbf{G_x})\sum_{G_p}P(\mathbf{G_p}|\mathbf{Ph})P(\mathbf{Ph})$$

equation 4

$$P(\mathbf{Ph}|\mathbf{G_p})=\sum_{G_{p2}}\sum_{G_{p3}}\ldots\sum_{G_{PN}}\sum_{Env}P(\mathbf{Ph}|\mathbf{G_p},\mathbf{G_{p2}},\mathbf{G_{p3}},\ldots,\mathbf{G_{PN}},\mathbf{Env})P(\mathbf{G_{p2}},\mathbf{G_{p3}},\ldots,\mathbf{G_{PN}},\mathbf{Env})$$

marker and (putative) trait locus genotypes are correlated (Box 1, equation 3; refs 15–17). If traits do not strongly predict underlying genotypes, that is, if $P(G_P|Ph)$ is small, linkage and LD mapping may have very low power or may not work at all. As an extreme example, one's genotype cannot be reliably determined by merely stepping on the bathroom scale! But even if this could be done, there is a widespread but invalid belief that because something can be mapped (that is, $P(G_P|Ph)$ is high), the causal predictive power of the genotype ($P(Ph|G_P)$; Fig. 1, blue arrow) will also be high. In fact, we have surprisingly little data on this latter topic, which requires extensive sampling from the general population, rather than patients. Note that the opposite can also be untrue—that is, if $P(Ph|G_P)$ is high it does not mean $P(G_P|Ph)$ will be high, as in genetically heterogeneous mendelian disorders such as retinitis pigmentosa[15,18]. It is important to note that when we speak of $P(Ph|G_P)$ in this context, we speak of the marginal mode of inheritance, which is only valid for consideration of singletons (Box 1, equation 4), and relatives will not have independent and identically distributed penetrances (even without assuming epistasis or gene-environment interactions) because the other genetic and environmental factors are also correlated among them! Similar arguments can be made about detectance, $P(G_P|Ph)$, which must always be a function of the ascertainment, something that is often overlooked in the literature when investigators make comparisons of power for different study designs[19].

Association studies must typically be carried out with a set of pre-selected SNPs, chosen according to whether they are expected to effectively represent the total existing variation. That is, in each region of the genome, the markers that are used are assumed to capture by LD the variation that exists at the unexamined sites in the region (Fig. 1, green arrow). Which and how many markers to use in mapping studies are relevant topics of debate. Most investigators seem to feel confident in using only a few sites per gene, such as 1 marker every 10, 30 or even 50 kb (ref. 20), and often seem quite casual about how these should be identified. It has even been suggested that typing only one SNP in every third or fourth gene may be sufficient to identify phenotypically active variants located anywhere in the entire genome[20]. We think the assumptions underlying such prognostications rosily underestimate the complexity of the problem, because, among other reasons, such estimates are often based on extrapolation from very small samples of data (which can lead to strong upward biases in the estimation of LD strength), or from predicted levels of LD (based on theory) that ignore or overlook their enormous stochastic variance[20–22]. And this does not begin to consider the more important influence of the typical weakness of genotype-phenotype relationships in complex traits[23,24].

Judging from most loci that have been sequenced in even modest sample sizes, we cannot reliably understand the haplotype (and hence LD) structure with only two or three (much less single) markers per gene[25–27]. The stochastic nature of LD is such that more rather than fewer markers (even if they are all intragenic) may be needed to yield high probability of detecting an aetiologic signal with achievable sample sizes, especially with globally distributed, common markers[28]. Whereas the details can be debated, even as many as 100,000 SNP markers will only yield about 1 per gene. Sorting through so many tests, while maintaining laboratory quality and statistical power, must be viewed as a daunting prospect, but may be ancillary to the real challenge, which is biological. At least, it is important to investigate these issues thoroughly with direct, detailed studies, rather than assuming that the problem is much simpler than it really is.

Kruglyak[10] used a simulation approach to arrive at a proposal that 500,000 markers would be required to identify genetic risk factors in a random genome scan. This had alarming repercussions, and invoked numerous counter-arguments, but even he may have provided an optimistic estimate, as his simulation was based on assumptions of great regularity in the behaviour of LD across the genome, and a simplified model of population history[12]. Some human populations have a comparatively simple structure and, perhaps, more LD conservation than those of large outbred populations. Even the consensus notion of how many SNPs may be needed to effectively mine the genomes of 'simple' populations seems to have risen to around 1,000,000 (ref. 29), an upward revision based on complexities that are being revealed as larger samples from a greater variety of populations are being studied. The distribution of LD has an enormous stochastic variance, and is typically highly skewed under even the best of conditions[19] such that while there certainly will exist some long regions of high LD in the genome[20,22], most pairs of SNPs will exhibit less LD than an extrapolation from the mean would predict.

There is also a risk of bias resulting from the fact that genomic regions identified and studied most intensely so far are those with relatively extensive LD, where genes are relatively easy to map. Moreover, there is a tacit implication that if two widely separated SNPs have strong LD between them, untyped intervening variation will also be in sufficient LD with the flanking SNPs that the latter will effectively capture the haplotype structure of the region. This may be true for some of the intervening variable sites, but is simply not a general rule. Of course, in some unusual populations, like the Saami[30], there may be extreme amounts of LD over long regions, but even in populations like that of Finland[31], or small parts thereof[32], there is little LD over long distances. In global populations with greater heterogeneity, the situation is certainly not going to be any simpler.

A more important point, more difficult to deal with and consequently given less attention, is that such predictions concern the mappability of one SNP against a map of other SNPs (Fig. 1, green arrow). However, the public-health challenge is to use a map of SNPs to map a gene whose variants have only a weak impact on a complex phenotype. Even completely sequencing all individuals will not solve this problem when Ph predicts G_P poorly, and in practice, few candid investigators who understand the situation believe that typing one SNP in one gene will enable reliable detection of LD with an adjacent candidate gene tens or even hundreds of kb away. But this is what is implied by the argument that only

30,000 SNPs are needed for a genome scan[20] (note that, to meaningfully distinguish true from false-positive signals, multiple testing demands that a greater marker density is required of a genome scan than for a study of a smaller number of candidate genes).

There is a palpable and understandable impatience to develop a complement of mapping SNPs as quickly and cheaply as possible. This is exemplified by the relatively expedient approach of identifying common coding variants[6,33] (cSNPs). The search for cSNPs can sometimes even be carried out by computer from EST databases[34,35], a kind of data-for-free approach (see, for example a study[36] reported on page 233 of this issue by Lee and colleagues). But is there really a free lunch? Proponents suggest that most alleles affecting common diseases will alter the coding sequence. If so, using cSNPs would require fewer markers to sort through—and accordingly fewer false positives[1]—as the causal variants might often be among the markers themselves. Yet EST databases are generated from a small number of samples, usually from healthy individuals, only haphazardly representing world populations. Some studies raise caution over relying on such a resource[6,33], which, once affordably available, might become established as the only available tool for many investigators.

Moreover, in this age of dramatic progress in understanding gene regulation, it seems strange to assume that most alleles affecting complex phenotypes will alter the protein structure. This is certainly not obvious in the case of metabolic traits that typically arise after decades of normal life. It is also inconsistent with what is known of the genetics that underlie similar phenotypes in other organisms[37–41]. Detailed studies now regularly identify non-coding variation influencing human disease as well[42–47]. Regulatory regions may be particularly relevant to chronic metabolic diseases that are consequent to inappropriate levels of enzyme expression, homeostatic epistasis, or response to environmental variation, leading to a gradual accumulation of damage over many years before reaching a critical threshold. When regulatory regions are involved, cSNPs serve as markers rather than causal sites themselves. But cSNPs may not be the best markers[10,13,20], especially if only relatively rare, nonsynonymous changes are used.

The fault, though brutal, lies not in our stats, but in ourselves, that we are underlink'd

If the real problem is not the relationship between two directly measured SNPs, but between one measured SNP and a phenotype that is at best a weak predictor of genotypes of an underlying SNP (if the disease is in fact 'genetic'), then how well can we expect to be able to identify these effects, and what type of resources would be needed to do it efficaciously? Each disease has its own genetic architecture that depends on human evolutionary history[23,24,48]. The pattern of variation is the product of past filtering by chance and selection—but filtering on phenotypes, not genotypes, and often inefficient filtering at that[49]. There is simply nothing in the basic process of evolution, not even strongly adaptive evolution, that forces G→Ph relationships to be strong or dominated by one or a small number of alleles or loci, and for complex traits we know the opposite is often true[50]. The human globin genes show this clearly. Subject to very strong natural selection imposed by malaria, the relatively simple phenotype 'resistance to malaria' (more clear-cut than 'susceptibility to heart disease') has been brought about by a large number of variants, in several genes, that vary within and among populations. We also know that as an almost unexceptionable statement, loci associated with simple, severe paediatric traits have many alleles that vary among populations and only a few of which are (relatively) frequent[13].

Late-onset chronic diseases—whose elimination through genetics is currently the supreme object of our affections—are much more complex by comparison. Present-day variation associated with late-onset diseases may have been selectively neutral in the past; at least, the disease-related effects of such variants would not have been detectable by natural selection because the afflicted would have already been able successfully to pass their genes to the next generation before the disease struck. Indeed, many of today's important chronic diseases were rare until a century or less ago. Variation in a gene associated with such a disease could thus have evolved neutrally, unless there were some early onset pleiotropic effects. Complex traits whose genetic basis has evolved neutrally are likely to have even 'noisier' genetic architecture than traits like malaria resistance. Multiple loci are almost always involved in complex traits, often with a plethora of risk alleles, and epistasis is likely to be common.

The most effective disease-related mapping and association studies are carried out in selective samples of individuals or families at high risk relative to the average risk in the population, and from populations with unusual histories. These individuals are in the tails of the risk distribution, which favours the ascertainment of alleles with large phenotypic effect, high penetrance and classically mendelian appearance in families. But such alleles are typically rare, and have low attributable risk (although they may be the most useful to target because their amelioration may have the greatest impact on quantity and/or quality of life of individual patients). Because most alleles have arisen only once in history, rare alleles are, as a rule, geographically localized, leading to additional heterogeneity among populations. Aetiology can be very heterogeneous, and it is no surprise to see the frequent elusiveness of mapping results[13,14]. For these and other reasons, regions identified by chronic disease mapping studies, and the candidate genes ultimately found in them, have often proved to have less direct public health impact than initially predicted from the rarefied samples in which they were first identified. For example, it is not clear how to interpret findings in which the (apparently) same variant has very different effects in different populations.

Investigators facing the failure of most plausible candidate genes to account for a high fraction of cases of a chronic disease, and frustrated by inconsistent or imprecise mapping efforts, with markers 1 to 10 cM (that is, approximately 1 to 10 million base pairs) apart, are investing resources and effort in the apparent belief that more markers and better statistical computing will make these traits meaningfully genetic. There will, of course, be exceptions, but we think genetic factors are not likely to explain these diseases in the usual causal sense. Instead, our frustrations are likely due to the biological realities, which involve the multiplicity of contributing factors (Fig. 1, blue arrows) that are typically unknown and largely ignored.

These include the all-too-important effects of environmental or non-genetic factors (Fig. 1, cream ovals), and their interactions with potential risk genotypes. It is obvious that environmental changes can produce dramatic phenotypic changes (note, for example, the height of doors to many old English pubs), and even that the environmental conditions in both pre- and post-natal development can strongly correlate with phenotypes that are expressed in later life[51]. These environmental conditions may be correlated among siblings, leading to inflated estimates of the 'genetic' component of these traits. As geneticists, we all too often treat the environment as a nuisance parameter, to be integrated out of our analyses, while criticizing traditional epidemiologists for treating genetic factors with similar disregard. And yet the traits we often focus on frequently have heritabilities of less than 50%, meaning most of the variation in the traits is not genetic in any simple sense. Geneticists focus little effort on controlling for potential environmental confounders, which may be more important than genetic factors in terms of having an impact on public health because they are more easily modified in many cases.

The best evidence for this is how much worse phenotypes related to chronic disease have become in recent times[52,53] (showing that they are resulting from concerted effects of genetics and the environment, which could be largely ameliorated by environmental change, as for example, most cases of type 2 diabetes). Particularly problematic targets of genetic intervention will be those common alleles with small average effect on risk that mappers seem most interested in finding. Because such alleles often interact with environmental exposure to agents such as diet and exercise that we know about, and modify in response to that knowledge, retrospective risk-estimates may be unreliable as the necessary basis for prospective risk estimates. For this reason it would seem imperative to control for environmental variation in the study design (rather than post-hoc statistical analysis) of genetic epidemiology investigations, if anything meaningful is to be learned[15].

Going straight for function?

Given these circumstances, some investigators argue for a different kind of genomics technology, called proteomics or functional genomics[54–58]. Directly characterizing the genes expressed in 'normalcy' and disease, it is said, will enable us to circumvent the need to sort through complex genetic variation to find the small minority of important sites. Some traits may be highly amenable to such approaches—for example, response to focused exposures to exogenous molecules like drugs or environmental toxins. But much of the enthusiasm for functional genomics may be illusory in ways similar to sequence-based approaches.

Gene expression levels are phenotypes that can be in 'cause' or 'effect' relationships to disease or exposure. For example, cancer cells will express tens or hundreds of genes associated with exuberant growth, aneuploidy, angiogenesis and so on, relative to adjacent normal tissue, but what fraction of these are 'causal'? Expression changes may serve as phenotypic risk factors, much as high cholesterol levels predict cardiovascular disease (for example, as markers for distinguishing different types of tumour), but the expression pattern itself may not be 'genetic' (that is, heritable). Many environmental factors affect gene expression. For example, levels of hormones vary dramatically with seasonal changes in the serum of the denizens of Svalbard[59,60] (a Norwegian archipelago in the far north). Regulation of blood pressure changes daily in response to many factors[61], levels of growth hormone and cortisol change in response to exercise[62], and nutritional factors during growth and development may affect metabolism and susceptibility to disease throughout later life[51]. We have evolved as homeostatic organisms that can respond to diverse environments in many ways, as evidenced by the fruits of our first explorations into the aetiological heterogeneity of chronic disease.

An array of challenges

Arrays of cDNA probes (cDNA microarrays) measure variation in mRNA level in some tissue at a given point in time. They generally do not detect changes in sequence variation (that is, structural changes) in genes, only differences in mRNA levels. Levels of mRNA (even if measured over time), are not even a complete measure of gene expression, let alone function. In fact, genetic variation that leads to modified protein structure may induce diverse expression responses in other genes that may be detectable, with such arrays, though the gene whose variation is influencing the expression profile (or the disease pathogenesis) would remain undetected. For example, coding changes in a transcription factor that are not detected in a cDNA expression array (meaning the mRNA levels of the transcription factor itself are unaltered) may result in a protein that binds anomalously and alters the expression of many other genes. The latter will appear to be different on the cDNA array, but those genes will not be causally affecting pathogenesis themselves.

Genetic factors affecting expression of a given protein often relate to other factors than the gene itself. In even the *lac* operon in *Escherichia coli*, expression levels of β-galactosidase can be mediated by sequence variants outside the coding region of *lacZ*, not to mention the environmental concentration of lactose or related compounds in the surrounding medium[63]. One can work back from a regulated gene with aberrant expression to find its regulator, but why would we expect the network of regulatory pathways to be less complex than the genetic heterogeneity with which we currently struggle?

Whereas many hypotheses may be ideally suited to testing by expression array, similar issues may challenge the interpretation of 'microarray' data as those that challenge sequence-based mapping and estimation of risk for complex disease. Another challenge is how to identify an appropriate candidate tissue[64] (or even cell types within the tissue) in which to compare expression in individuals with different phenotypes. And, once identified, how to obtain samples (for example, schizophrenics—and the matched controls—may be reluctant to offer brain biopsies). This contrasts with the ability to use almost any tissue to identify inherited genetic variation. How well can we identify appropriate tissues when it comes to complex chronic diseases? The answer is not obvious. Whereas analysis by microarray may be well suited to some situations, the statistical and methodological issues that are apparent in a genome screen do not convincingly seem to favour it as a general approach over a genetic one. Of course the ability to use array technology to measure the expression levels of 100,000 genes over time means 100,000 phenotypes could be studied in the same genome scan, some of which might be under tractable genetic influence. So focusing on the genetics of the expression level itself, rather than disease, may yield interesting discoveries related to basic biology and perhaps pathobiology, if not aetiology.

"*Prudens quæstio dimidium scientiæ*" (To know what to ask is already to know half)—Aristotle

SNPs are more numerous but individually less informative than microsatellites (which are more polymorphic), and cannot rescue a question not properly posed (that is, using an inappropriate experimental design). We have discussed mapping and association approaches that are being developed with enthusiasm, and at great expense, justified by promises of major advances in chronic disease[13,14]. Even for a properly framed question and an appropriately designed ascertainment scheme, the popular methods work well only when one can reliably predict disease-locus genotypes from phenotypes.

The filling of the journals with 'positive' results certainly gives the appearance of a relentless juggernaut of successes, and it may seem incredible that we could raise the level of skepticism we have. But most of these results, including those implicating *BRCA1*, *BRCA2* (in breast cancer), and *PSEN1*, *PSEN2* (encoding the presenilins, which have variants that predispose to Alzheimer disease) and the like, have been obtained through linkage analysis of multiplex pedigrees as opposed to case-control LD studies. Even ignoring the elusive positives from case-control LD studies, are we being misled by the 'successful' mapping 'hits', from the point of view of public health?

Ewens[65] once pointed out that the medical system in developed nations is designed to screen and report on hundreds of diseases and will of course report on those that are rare. If these are chosen selectively for study, or results of their study reported selectively, a misleading impression can result that common diseases are commonly 'genetic'. This leads to our suggestion that it might be instructive to invert the fundamental question, and ask: how many instances of disease, or families, or samples do we have to search before we find an example of a SNP with high attributable

risk, or that at least is sufficiently predicted from phenotypes to be mappable? If the number is large, in other words, if the preponderance of these efforts are going to yield relatively minor, negative, uninterpretable or irreproducible results, the approach—or the concepts on which it is based—is inefficient or inappropriate.

Common alleles of 'strong effect' do of course exist, and have been identified, often through candidate gene studies. Examples in the peer-reviewed literature in which the data have been subjected to scrutiny include *DCP1*, sickle cell anaemia and *HBB*, psoriasis in the MHC region[66], and the well-documented *APOE* protein isoform alleles. The association between *APOE* ε4 alleles and Alzheimer disease in populations of European descent appears to be replicable, although in African Americans, the 'risk' genotypes, which are more frequent than in Europeans, have inconsistent (and sometimes even negative) association with the disorder[67–69]. This highlights the fickleness of even the most celebrated examples of complex disease mapping success, and demonstrates the importance of population choice in the best of cases. The same scenario arises regularly in other cases, such as the intensively examined association of calpain-10 with type 2 diabetes reported by Bell and colleagues on page 163 of this issue[47] (and that was, incidentally, first found by linkage analysis rather than association); how well it will stand up to future tests and extrapolation remains to be seen, but even those wearing the most rose-coloured of glasses would have to acknowledge that the impact is hugely variable as a function of genetic, population and environmental context.

"I found one! I found one!"

We do not contest that genome scanning for LD can work when there is a strong difference in the frequency of some risk genotype(s) between cases and controls. Nikali *et al.*[70] were able to map the gene responsible for a rare recessive ataxia in a Finnish cohort using an LD-based genome scan with microsatellite markers using only four affected individuals, demonstrating that it can, of course, work. Similarly, the poster boy of the SNP proponents wears a T-shirt advertising *APOE* ε4 and Alzheimer disease; this association is cited as 'proof-of-principle' for SNP-based association mapping. By using a data set in which *APOE* ε4 was known to be over-represented in cases compared with controls, it has been possible to take a set of SNPs that are in LD with the risk allele and identify the association without studying *APOE* ε4 itself[71,72]. This case, stacked in favour of positive findings, shows that neither coding SNPs, nor most SNPs, nor common SNPs, nor nearby SNPs, nor single SNPs, nor a SNP in a neighbouring gene, nor even simple haplotypes can be relied upon, *a priori*, as a single, casually designed source of LD markers. What this does show is the not-surprising fact that when there is an association with a single risk allele, one can identify this by using multiple markers which are in LD with this single risk allele[14]. How easily would genes such as *BRCA1*, *BRCA2*, *PSEN1*, *PSEN2* and *ABCA4*, which have many alleles each of which increases risk of some common disease, be mapped by scrutinizing a set of common SNPs in a sample from a large cosmopolitan population[13,14]?

No one can deny that disease pathways have been identified by genetic epidemiology studies, making contributions to our biological knowledge. So far, however, it is lifestyle changes that have made the most impact on reduced (or increased!) incidence of chronic disease[52,53,73]. In those instances in which common variants affecting common disease really do exist, it is important to find them. But we regularly hear grander promises, and it is at least fair to ask whether scaling up current genetic approaches, which have been likened to a search for a needle in a needle stack (a great many individually modest, effects), would be the wisest investment when a major justification is that nothing else has worked so far.

To some this is an inapt question. 'New' genes may identify previously unknown biochemical pathways that may lead to therapeutic pharmaceutical improvements (even if the latter are not specifically 'genetic'). Other investigators argue that the identification of even weakly predictive screening tools is sufficient justification, or hope that a worthwhile fraction of complex diseases will have common variants with individually modest effect but substantial attributable risk[1–3,7]. The level of merit in these arguments is difficult to evaluate, and the track record is unclear at best; but the arguments undeniably have aspects of rationalizing results that are not living up to expectations, sometimes based on unrealistically optimistic models[20,21,74], and sometimes dominated by wishful thinking. At the least, we should undertake the effort to assess the underlying assertions—including of course those that we advance here—about the degree to which common alleles with usefully important effects exist for chronic disease; such asssessment should use focused and systematic approaches that take the biological realities seriously.

We know these views run against momentum and heavily vested interests. Those interests are promoted by reference to their successes (sometimes overstated, and often argued first—or only—through the business or journalistic press rather than peer-reviewed articles[4]). We expect criticism from such quarters, but readers should be aware of the less-publicized failures and inefficiencies. What per cent of complex disease gene-mapping projects whose grant proposals promised 80% power have actually been successful in identifying the disease genes assumed to be in the data? Application of most current mapping strategies could arguably be more effective if focused on traits that clearly have a genetic basis, such as frank paediatric disorders (which represent a human version of mouse knockouts, where one can assess the effects of the complete absence of a gene product), or on candidate genes, where the genetic or population payoff might be higher, or subsets of complex disease, in samples from appropriately chosen populations where LD may be high and aetiology more marked and homogeneous (or perhaps most importantly

through reductions in *environmental* heterogeneity). Even if such traits are individually rare, they are numerous[75], and we know genetic methods work to unravel their aetiologies.

Resistance to genetic reductionism is not new[76], and we know that, by expressing these views (some might describe them as heresies), we risk being seen as stereotypic nay-sayers. However, ours is not an argument against genetics, but for a revised genetics that interfaces more intimately with biology. Biological traits have evolved by noise-tolerant evolutionary mechanisms, and a trait that doesn't manifest until long after the reproductive lifespan of most humans throughout history is unlikely to be genetic in the traditional, deterministic sense of the term. Most genetic studies that focus on humans are designed, in effect, to mimic Mendel's choice of experimental system, with only two or three frequent states with strongly different effects. That certainly enables us to characterize some of the high-penetrance tail of distribution of the allelic effects, but as noted above these may usually be rather rare. But inflated claims based on this approach can divert attention from the critical issue of how to deal with complexity on its own terms, and fuel false hopes for simple answers to complex questions. The problems faced in treating complex diseases as if they were Mendel's peas show, without invoking the term in its faddish sense, that 'complexity' is a subject that needs its own operating framework, a new twenty-first rather than nineteenth—or even twentieth—century genetics.

Acknowledgements

Support to J.D.T. is acknowledged from a Hitchings-Elion Fellowship from the Burroughs-Wellcome Fund, and to K.M.W. from NIH grant HL 58239. The opinions in this article are the personal views of the authors, and do not necessarily reflect the views or policies of the funders.

Received 25 February; accepted 11 May 2000.

1. Brookes, A. The essence of SNPs. *Gene* **234**, 177–186 (1999).
2. Collins, F.S. The human genome project and the future of medicine. *Ann. N. Y. Acad. Sci.* **882**, 42–55 (1999).
3. Collins, F.S. Genetics: an explosion of knowledge is transforming clinical practice. *Geriatrics* **54**, 41–47 (1999).
4. Russo, E. Bypassing peer review. *The Scientist* **14**, 1–12 (2000).
5. Sander, C. Genomic medicine and the future of health care. *Science* **287**, 1977–1978 (2000).
6. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
7. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
8. Collins, F., Brooks, L. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
9. Pastinen, T. *et al.* Array-based multiplex analysis of candidate genes reveals two independent and additive genetic risk factors for myocardial infarction in the Finnish population.I. *Hum. Mol. Genet.* **7**, 1453–1462 (1998).
10. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
11. Wang, D. & Lander, E. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
12. Zöllner, S. & von Haeseler, A. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **66**, 615–628 (2000).
13. Terwilliger, J.D. & Weiss, K.M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotech.* **9**, 578–594 (1998).
14. Terwilliger, J.D. & Goring, H.H. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum. Biol.* **72**, 63–132 (2000).
15. Terwilliger, J.D. A likelihood-based extended admixture model of oligogenic inheritance in "model-based" or "model-free" analysis. *Eur. J. Hum. Genet.* (in press).
16. Göring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors. III. Marker loci and their map as nuisance parameters. *Am. J. Hum. Genet.* **66**, 1298–1309 (2000).
17. Göring, H.H. & Terwilliger, J.D. Linkage analysis in the presence of errors. I V. 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.* **66**, 1310–1327 (2000).
18. Heckenlively, J. & Daiger, S. Heredity and retinal and choroidal degenerations. in *Emory and Rimoin's Principles and Practice of Medical Genetics* (eds Rimoin, D., Connor, J. & Pyeritz, R.) 2555–2576 (Churchill-Livingstone, Edinburgh, 1996).
19. Terwilliger, J.D. On the resolution and feasibility of genome scanning approaches to unraveling the genetic components of multifactorial phenotypes.I. in *Genetic Dissection of Complex Phenotypes: Challenges for the Next Millennium* (eds Rao, D.C. & Province, M.A.) (Academic, New York, 2000).
20. Collins, A., Lonjou, C. & Morton, N.E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
21. Schork, N.J., Cardon, L.R. & Xu, X. The future of genetic epidemiology. *Trends Genet.* **14**, 266–272 (1998).
22. Ott, J. Predicting the range of linkage disequilibrium. *Proc. Natl Acad. USA* **97**, 2–3 (2000).
23. Weiss, K.M. *Genetic Variation and Human Disease: Principles and Evolutionary Approaches* (Cambridge University Press, Cambridge, 1999).
24. Weiss, K. Is there a paradigm shift in genetics? Lessons from the study of human diseases. *Mol. Phylogenet. Evol.* **5**, 259–265 (1996).
25. Clark, A.G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
26. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
27. Templeton, A.R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).
28. Sing, C. *et al.* Genotype-phenotype studies based on the full DNA sequence of the Apo E gene demonstrate the challenge we face in the assignment of function to a particular DNA polymorphism. *Am. J. Hum. Genet.* **65**, A14 (1999).
29. Roberts, L. Human genome research. SNP mappers confront reality and find it daunting. *Science* **287**, 1898–1899 (2000).
30. Terwilliger, J.D., 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.* **48**, 148–154 (1998).
31. Eaves, I.A. *et al.* The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes *Nature Genet.* **25**, 320–323 (2000).
32. Varilo, T. *et al.* Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur. J. Hum. Genet.* **8**, 604–612 (2000).
33. Halushka, M.K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
34. Garg, K., Green, P. & Nickerson, D. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res.* **9**, 1087–1092 (1999).
35. Picoult-Newberg, L. *et al.* Mining SNPs from EST databases. *Genome Res.* **9**, 167–174 (1999)-
36. Irizarry, K. *et al.* Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nature Genet.* **26**, 233–236 (2000).
37. Beckwith, J. & Zipser, D. (eds) *The Lactose Operon* (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1970).
38. Long, A., Lyman, R., Langley, C. & Mackay, T.F. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in Drosophila melanogaster. *Genetics* **149**, 999–1017 (1998).
39. Mackay, T.F. The nature of quantitative genetic variation revisited: lessons from Drosophila bristles. *Bioessays* **18**, 113–121 (1996).
40. Stam, L. & Laurie, C. Molecular dissection of a major gene effect on a quantitative trait: the level of alcohol dehydrogenase expression in Drosophila melanogaster. *Genetics* **144**, 1559–1564 (1996).
41. Hendrich, Z. & Willard, H. Epigenetic regulation of gene expression—the effect of altered chromatic structure from yeast to mammals. *Hum. Mol. Genet.* **4**, 1765–1777 (1995).
42. Artiga, M. *et al.* Risk for Alzheimer's disease correlates with transcriptional activity of the *APOE* gene. *Hum. Mol. Genet.* **7**, 1887–1892 (1998).
43. Hall, S. *et al.* A common mutation in the lipoprotein lipase gene promoter, -93T/G, is associated with lower plasma triglyceride levels and increased promoter activity in vitro. *Arteriorscler. Thromb. Vasc. Biol.* **17**, 1969–1976 (1997).
44. Gragnoli, C. *et al.* Maturity-onset diabetes of the young due to a mutation in the hepatocyte nuclear factor-4 α binding site in the promoter of the hepatocyte nuclear factor-1 α gene. *Diabetes* **46**, 1648–1651 (1997).
45. Scholtz, C. *et al.* Mutation -59C→T in repeat 2 of the LDL receptor promoter: reduction in transcriptional activity and possible allelic interaction in a South African family with familial hypercholesterolemia. *Hum. Mol. Genet.* **8**, 2025–2030 (1999).
46. Steel, C. Cancer of the breast and female reproductive tract. in *Emory and Rimoin's Principles and Practice of Medical Genetics* (eds Rimoin, D., Connor, J. & Pyertiz, R.) 1501–1524 (Churchill-Livingstone, Edinburgh, 1996).
47. Horikawa, Y. *et al.* Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175 (2000).
48. Weiss, K. In search of human variation. *Genome Res.* **8**, 691–697 (1998).
49. Schlichting, C. & Pigliucci, M. *Phenotypic Evolution: A Reaction Norm Perspective* (Sinauer Associates, Sunderland, 1998).
50. Weiss, K. & Fullerton, S. Phenogenetic drift and the evolution of genotype-phenotype relationships. *Theor. Popul. Biol.* (in press).
51. Waterland, R. & Garza, C. Potential mechanisms of metabolic imprinting that lead to chronic disease. *Am. J. Clin. Nutr.* **69**, 179–197 (1999).
52. Trowell, H. & Burkitt, D. *Western Diseases: Their Emergence & Prevention* (Harvard University Press, Cambridge, Massachusetts, 1981).
53. Shephard, R. & Rode, A. *The Health Consequences of "Modernization"* (Cambridge University Press, Cambridge, UK, 1996).
54. Anderson, N.L. & Anderson, N.G. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**, 1853–1861 (1998).
55. Brent, R. Functional genomics: Learning to think about gene expression data. *Curr. Biol.* **9**, 338–341 (1999).
56. Brent, R. Genomic biology. *Cell* **100**, 169–183 (2000).
57. Schena, M. *et al.* Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**, 301–306 (1998).
58. Woychik, R., Klebig, M., Jusice, M., Magnusson, T. & Avrer, E. Functional genomics in the post-genome era. *Mut. Res.* **400**, 3–14 (1998).
59. Bojko, E. & Larsen, T. Changes in the serum lipid proflie in man during 24 months Arctic residence. *Int. J. Circumpolar Health* **58**, 170–175 (1999).
60. Bojko, E. Metabolical changes induced by adaptation to circumpolar conditions in Spitzbergen. *Int. J. Circumpolar Health* **56**, 134–141 (1997).
61. Rowell, L.B. *Human Cardiovascular Control* (Oxford University Press, Oxford, 1993).
62. Hatfield, F. & Platz, T. *Hardcore Bodybuilding: a Scientific Approach* (Contemporary Books, Chicago, 1993).
63. Schlief, R. *Genetics & Molecular Biology* (Addison-Wesley, Reading, Massachusetts, 1986).
64. Cole, K.A., Krizman, D.B. & Emmert-Buck, M.R. The genetics of cancer—a 3D model. *Nature Genet.* **21**, 38–41 (2000).
65. Ewens, W.J. Tay-Sachs disease and theoretical population genetics. *Am. J. Hum. Genet.* **30**, 328–329 (1978).
66. Trembath, R. *et al.* Identification of a major susceptibility locuson chromosome 6p and evidence for further disease loci revealed by a two stage genome-wide search in psoriasis. *Hum. Mol. Genet.* **6**, 813–820 (1997).
67. Tang, M.X. *et al.* Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. *Am. J. Hum. Genet.* **58**, 574–584 (1996).
68. Farrer, L.A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**, 1349–1356 (1997).
69. Tang, M. *et al.* The APOE-epsilon4 allele and the risk of Alzheimer disease among African-Americans, Whites, and Hispanics. *JAMA* **2789**, 751–755 (1998).
70. Nikali, K. *et al.* Random search for shared chromosomal regions in four affected individuals. *Am. J. Hum. Genet.* **56**, 1088–1095 (1995).
71. Martin, E.R. *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in alzheimer disease. *Am. J. Hum. Genet.* **67**, 383–394 (2000).
72. Martin, E.R. *et al.* Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* **63**, 7–12 (2000).
73. Bjerregard, P. & Young, T. *The Circumpolar Inuit: Health of a Population in Transition* (Munksgaard, Copenhagen, 1998).
74. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
75. McKusick, V.A. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders* (Johns Hopkins University Press, Baltimore, Maryland, 1998).
76. Sarkar, S. *Genetics and Reductionism* (Cambridge University Press, Cambridge, UK, 1998).