Linkage disequilibrium mapping of complex disease: fantasy or reality?

Joseph D Terwilliger* and Kenneth M Weiss†

In the past year, data about the level and nature of linkage disequilibrium between alleles of tightly linked SNPs have started to become available. Furthermore, increasing evidence of allelic heterogeneity at the loci predisposing to complex disease has been observed, which has lead to initial attempts to develop methods of linkage disequilibrium detection allowing for this difficulty. It has also become more obvious that we will need to think carefully about the types of populations we need to analyze in an attempt to identify these elusive genes, and it is becoming clear that we need to carefully re-evaluate the prognosis of the current paradigm with regard to its robustness to the types of problems that are likely to exist.

Addresses
*Columbia University, Department of Psychiatry and Columbia Genome Center, 60 Haven Avenue #15-C, New York, NY 10032, USA; e-mail: joseph.terwilliger@columbia.edu
†Department of Anthropology, Pennsylvania State University, 409 Carpenter, University Park, PA 16802, USA; e-mail: Kmw4@psu.edu
Correspondence: Joseph D Terwilliger

Current Opinion in Biotechnology 1998, **9**:578–594

http://biomednet.com/elecref/0958166900900578

 ISSN 0958-1669

Abbreviations

+	not carrying disease-predisposing alleles
D	disease-predisposing allele
LD	linkage disequilibrium
SNP	single nucleotide polymorphism
TDT	transmission/disequilibrium test

Introduction

The Chinese philosopher Confucius was asked by one of his disciples how to distinguish a good man from a bad man, to which Confucius replied (Lun Yu 15:27), "What all men praise, examine critically; what all men condemn, examine critically" [1]. The use of allelic association mapping in the search for complex-disease-predisposing genes is an idea that has received much praise in the literature over the past few years, despite a lack of convincing supportive evidence. In surveying the state of this research area, perhaps the most important questions to examine are those that have not been looked into critically often enough in the literature. There is a widespread belief that somehow the advent of a genome- spanning map of single nucleotide polymorphisms (SNPs) will provide some sort of panacea for the woes that have been plaguing those of us trying to unravel the complex etiology of common genetic disease [2•]. Certainly this is one more tool in our arsenal of weapons, but we must be careful to critically examine the assumptions that underlie the prognosis for this method's success before jumping on the bandwagon in the search for a simple and guaranteed solution. As Confucius put it, "The gentleman agrees with others without being an echo. The small man echoes others without being in agreement." (Lun Yu 13:23) [1].

Association analysis in humans has been performed successfully for fine mapping of a large number of genes that have large effects on rare phenotypes that segregate in pedigrees. Most of these disease-predisposing loci had been previously mapped with linkage analysis by following the segregation of the disease in pedigrees. There are a large number of diseases that are far more common, yet tend to occur more frequently among relatives of affected individuals than in the general population and have substantial heritability, yet there is no clear pattern of segregation in families. Because there is a clear genetic component to these diseases, it is widely believed that allelic association and linkage analysis methods will be able to identify the genes underlying these complex common traits as well. The difficulty is that individuals with a given disease may be affected for completely different genetic reasons. The main difficulty is that the effect of any allele on the risk for chronic disease is typically weak — otherwise one would observe clear patterns of phenotypic segregation in large pedigrees.

First let us describe what the terms allelic association and linkage disequilibrium (LD) refer to. If there are two tightly linked loci, with two alleles each (A,a at locus 1; B,b at locus 2), then there are four possible combinations of alleles that could exist on the same chromosome, A_B, A_b, a_B and a_b. If allele A has frequency p_A, and allele B has frequency p_B, then the haplotype A_B would have frequency $p_A p_B$, for example, in the absence of linkage disequilibrium (i.e. the alleles occur independently on haplotypes). If alleles A and B are associated, the frequency of haplotype A_B would be $p_A p_B + \delta$, where δ is a measure of the strength of LD between the two loci. If allele B at locus 2 predisposes to some disease phenotype, then if one ascertains a sample of affected individuals (cases) from the population, and a sample of unaffected individuals (controls), then allele A would be found more frequently in cases than controls. In other words, there would be an association between allele A and the disease phenotype. In practice, one can test a large number of marker loci throughout the genome, or a set of polymorphisms in or around a candidate gene, in the hope that one of these marker loci would be close enough to a disease locus that some marker allele might be associated with the disease allele. This is the basis of association and LD mapping, which has been shown to work well in the case of

simple disease in populations where there is likely to have been only one disease-predisposing allele at this locus (e.g. [3]).

The hope of many clinically minded individuals is that association analysis methods will lead to early diagnosis and treatment of chronic common disease with greater accuracy — through the identification of genetic risk factors. For many of these phenotypes, environmental risk factors (e.g. smoking and lung cancer) or phenotypic risk factors (high serum cholesterol and coronary heart disease) are already known to be valuable predictors of disease outcome. Estimating risk for even some of these more proximate risk factors can be very difficult, but in principle, at least, if genetically at-risk individuals can be identified, preventative measures might be taken before they have suffered environmentally induced damage. Can genotypic data help improve the ability to predict or treat these diseases? Are there genetic risk factors with equivalently strong effects? Can we hope to develop the magic pill that will circumvent the need for diet and exercise in the treatment of obesity? Different investigators are interested in answering each of those questions through the use of association and linkage analyses to identify the potential risk-increasing genotypes. The purpose of this review will be to examine the current state of the science to see what empirical and theoretical results have been made towards answering these questions. It is almost certain that association and linkage mapping will identify some alleles that have some etiological effect on some chronic disease; however, there remains much confusion about when and how this will work.

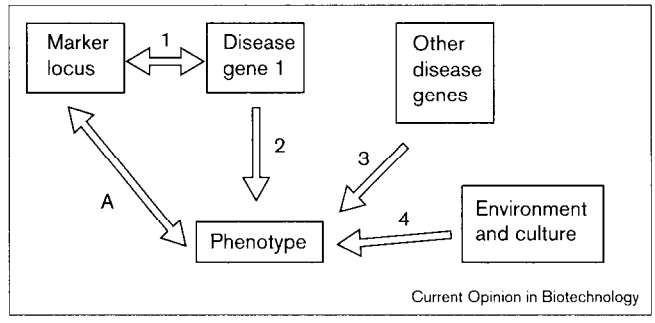
Complex disease

In order to evaluate the literature and the advances of the past year in association mapping, it is important to establish a model for the causation of complex disease. Consider the causal relationships in Figure 1. In the search for complex disease-predisposing genes, the main objective is to identify a marker locus that is correlated (Path 1) with a given disease gene (Disease gene 1), which itself is somehow influencing (Path 2) the phenotype. Because we do not know the identity of the disease gene, however, all we are able to identify are correlations along the Path A, which is some convolution of the correlations on Path 1 and Path 2. If the correlations along either Path 1 or Path 2 are negligible, the null hypothesis of no correlation between marker locus and phenotype will not be rejected.

We also know that other genetic factors (Path 3) and a myriad of environmental and cultural factors (Path 4) influence the phenotype. These various etiological factors will almost certainly interact with each other in a variety of complex and often intractable ways. The goal in setting up any linkage or association study to map disease-predisposing genes is to minimize the effects of Path 3 (genetic factors), Path 4 (environmental and cultural factors), and the interactive pathways, while maximizing the correlations on Path 1 and Path 2. Researchers usually attempt to minimize the effects of Path 3 and Path 4 by selecting homogeneous genetic and cultural isolates for study. In reality, these populations are often not as homogeneous as one might think, but nevertheless an attempt is made to decrease the residual genetic variation unrelated to disease through choice of study samples. This ascertainment process may sometimes increase the relative effects of one or few alleles in one or few genes as well, through the reduction in genetic heterogeneity, which would increase the correlation along Path 2. A means to increase the correlations along Path 1 is to increase the number of marker loci studied. When more loci are studied there is an increased probability that one of them will be strongly correlated with Disease gene 1, though of course this will also increase the type I error (false positive rate) due to the increased number of statistical tests. Additionally, Path 2 is often strengthened through the examination of very specific or idiosyncratic phenotypes, such as early-onset forms of diseases, or unusual high-penetrance variants (i.e. anyone carrying the disease allele is very highly likely to have the disease) that might be more easily mapped.

There are a number of situations in which certain quantitative phenotypes are also known to be risk factors for a common disease phenotype, for example, high serum cholesterol is a phenotypic risk factor for coronary heart disease. In this model, genetic and environmental factors may influence the quantitative phenotype more directly, in which case the genetic basis of that phenotype might be easier to dissect, as it may be more immediately under genetic control; however, even these traits turn out to be genetically quite complex.

Figure 1

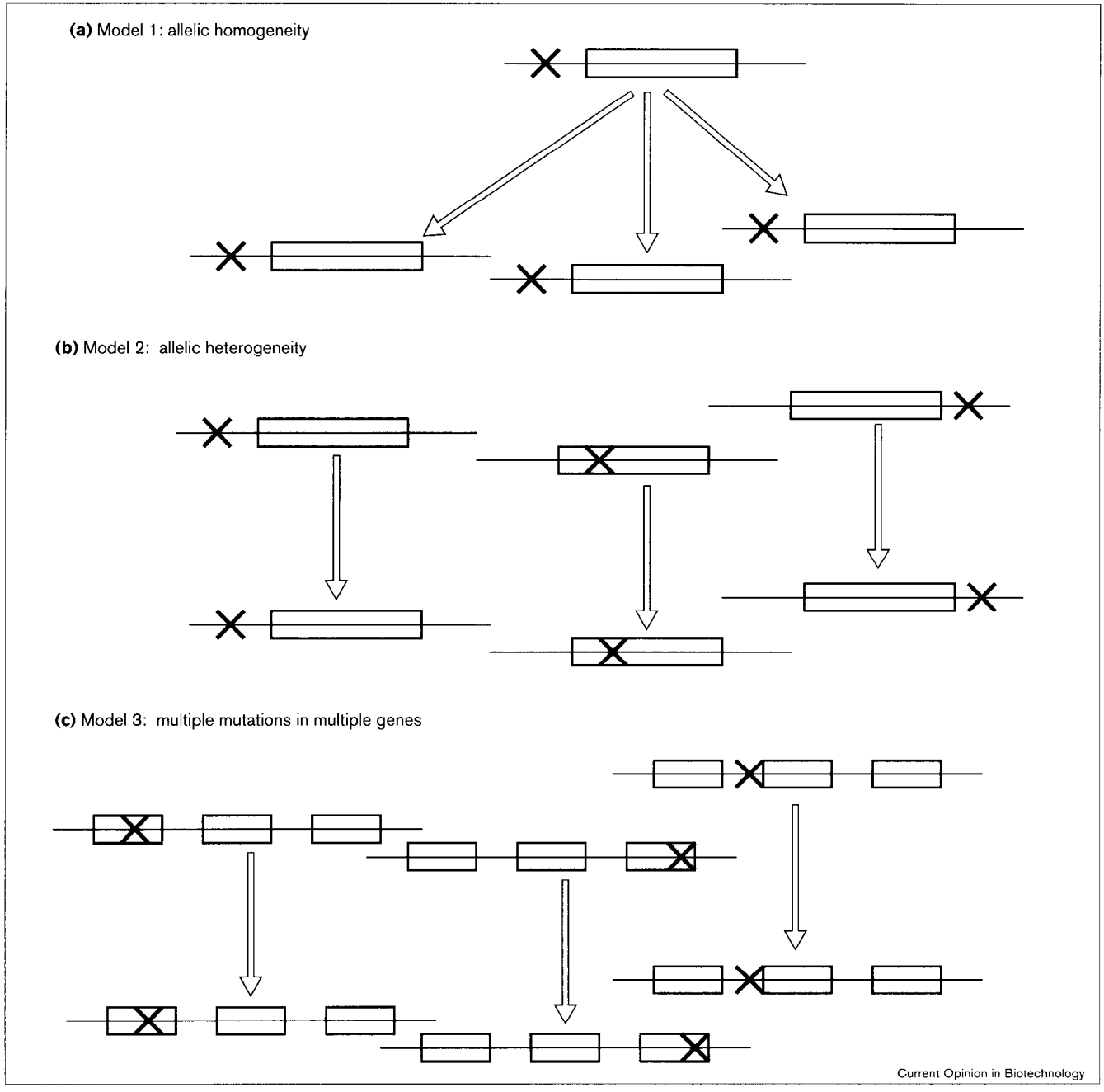


A simplified model of the etiological factors predisposing to complex phenotypes is diagramed. Each factor shown is assumed to have some predisposing role, and there are numerous potential interactions between them (not shown). In linkage or association analysis, we are testing for correlations between marker locus genotypes and complex phenotypes (Path A), which is a secondary correlation, rather than the true genotype–phenotype correlation shown as Path 2. The success of association mapping for complex disease depends, therefore, both on the correlation between marker loci and the disease locus genotypes (Path 1) and the correlation between genotype and phenotype (Path 2). A convolution of these two factors is what association and linkage mapping attempt to use to identify disease-predisposing loci.

Allelic architecture of complex disease

The next issue to consider in complex disease is how much allelic complexity there will be at any given disease-predisposing locus. There are a number of possible models of allelic architecture, as outlined in Figure 2. Risch and Merikangas [4] proposed that the future of complex-disease gene mapping will most probably be based on association mapping. Although their deductions are not implausible, it is the premise from which the deductions were made that must be examined critically. The allelic architecture they assumed (Model 1 in Figure 2) is that in a given gene there is one-and-only-one disease-predisposing allele (D) and

Figure 2



Three simple models for the allelic complexity of genetic disease are shown. **(a)** In Model 1, all disease-predisposing alleles at a given locus are identical by descent in the population – having derived from some common ancestor. In this situation, there is expected to be a conserved haplotype around the disease allele, which is shared by all carriers in the population many generations later. **(b)** Model 2 shows the case of allelic heterogeneity, in which multiple different allelic variants can each predispose to the phenotype. Thus among individuals with one of these 'D' alleles, there will be an assortment of haplotype backgrounds. The more heterogeneity, the less LD. **(c)** Model 3 shows the situation for multiple 'D' alleles in different genes. These genes may be linked (as shown) or unlinked.

Table 1

A selection of disease-predisposing loci with multiple different alleles predisposing to disease – mutation analyses presented in American Journal of Human Genetics volumes 60, 61, 62 (1997–8).

Gene symbol	Disease	Reference
PTCH	Nevoid basal cell carcinoma	[105]
ATP7A	Menkes disease	[106]
MAT1A	Hypermethioninemia	[107]
COL17A1	Benign epidermolysis bullosa	[108]
TGM1	Autosomal recessive congenital icthyosis	[33]
BRCA1	Breast cancer	[78]
BRCA2	Breast cancer	[78]
CFTR	Cystic fibrosis	[109]
Presinilin 1	Alzheimer's disease	[110]
TIMP3	Sorsby fundus dystrophy	[111]
WRN	Werner syndrome	[112]
Cystatin B	Progressive myoclonus epilepsy	[113]
TCS	Treacher–Collins syndrome	[114]
COL5A1	Ehlers-Danlos syndromes I and II	[115]
Various	Nonsyndromic hearing impairment	[16•]
PEX	X-Linked hypophosphatemic rickets	[116]
HEXA	Tay–Sachs disease	[117]
CANP3	Limb-girdle muscular dystrophy type 2A	[118]
PBG deaminase	Acute intermittent porphyria	[119]
OCRL1	Lowe oculocerebrorenal syndrome	[120]
PKD1	Polycystic kidney disease 1	[121]
SMN^T	Spinal muscular atrophy	[28]
ALK-1	Hereditary hemorrhagic telangiectasia type 2	[122]
RB1	Retinoblastoma	[123]
CLN3	Batten disease	[124]
ATP7B	Wilson disease	[125]
EXT1 and EXT2	Hereditary multiple exostoses	[126]
PKD2	Polycystic kidney disease	[127]
RPGR	Retinitis pigmentosa	[128]
COL7A1	Dystrophic epidermolysis bullosa	[129]
Myosin VIIA	Usher syndrome 1B	[130]
PKD1	Renal cystic disease in tuberous sclerosis	[131]
FBPase	Fructose 1,6 diphosphate deficienty	[132]
G4.5	Infantile dilated cardiomyopathies	[41]
MMAC1	Early-onset breast cancer	[133]
G4.5	Barth syndrome	[43]
TIGR	Primary open angle glaucoma	[134]
FAA	Fanconi anemia	[135]
PTEN	Breast cancer, cowden disease, juvenile polyposis	[136]
PDX1	Lactic acidosis	[137]
NAGLU	Sanfilippo syndrome type B	[138]
CSB (ERCC6)	Cockayne syndrome	[139]
ATM	Ataxia-telangiectasia	[46]
MEN1	Multiple endocrine neoplasia type I	[140]
HMG-CoA Lyase	HMG CoA lyase deficiency	[141]
alpha-TTP	Ataxia with isolated vitamin E deficiency	[45]
COMP	Pseudoachrondroplasia	[142]
CYP1B1	Primary congenital glaucoma	[143]
Arylsulfatase E	X-Linked chondrodysplasia punctata	[144]
UGT1A1	Criler-Najjar syndrome type I	[39]
HPS	Hermansky-Pudlak syndrome	[145]
RYR1	Malignant hyperthermia	[48]
HGO	Alkaptonuria	[146]
PYGL	Glycogenosis type VI (Hers disease)	[147]
GJB2	Autosomal recessive hearing loss	[148]
OA1	X-Linked ocular albinism	[149]
WT1	Isolated diffuse mesangial sclerosis	[47]
Btk	X-linked agammalobulinemia	[150]
PCBD	Hyperphenylalaninemia	[151]
Na-K-2Cl Cotransporter	Antenatal Bartter syndrome	[152]
Ferrochelatase	Erthropoetic protoporphyria	[153]
UBE3A	Angelman syndrome	[154]
JAG1	Alagille syndrome	[155]
TWIST/FGFR	Saethre–Chotzen syndrome	[156]

that said allele has an identical etiological effect in all individuals, related or not, and that said allele is being tested directly in the association analysis. For the rare monogenic recessive diseases of the Finnish disease heritage [5] this has provided a reasonable first-order approximation to the real situation. The reason is that a small founding population is unlikely to carry more than one allele for a rare disorder, so that ascertaining descendants by inheriting two defective alleles is essentially ascertaining that founding haplotype. This is, however, the only model in which extended haplotypes would be expected to be shared by all disease allele carriers, though even for these diseases there are cases of allelic heterogeneity and gene conversion or recurrent mutation disrupting those shared haplotypes (similar results have been found for Hirschsprung's disease in the Amish [6] and recessive diseases in the French Canadians [7,8]). Kaplan *et al.* [9] have described additional problems with LD mapping in complex populations.

The next simplest model (Model 2 in Figure 2) would be that of multiple unique, but functionally equivalent alleles in the same gene, such as appears to be the case on the macroscopic level for BRCA1, retinitis pigmentosa, cystic fibrosis, and many others (Table 1). A more complex and more realistic variation of Model 2 is that different alleles in the same gene might have different quantitative effects on the phenotype, as appears to be the case for most of these genotypes when they are examined at sufficiently microscopic scale (see [10,11,12•,13]). Bale [14] has pointed out that "variable expression is the rule rather than the exception."

A still more complicated model (Model 3 in Figure 2) that is closer to reality is that of multiple disease-predisposing alleles in multiple genes. Very often, functionally related genes are linked to each other, and a single linkage signal to a region of a chromosome may actually be the result of disease-predisposing alleles in different linked genes in different pedigrees. This was the case for X-linked retinitis pigmentosa [15], where the individual signals were able to be separated, and the existence of several genetic defects in the same chromosomal region was proven through linkage analysis alone. Had the effect of the disease alleles been to increase the susceptibility to the disease only 2–5-fold rather than being fully penetrant recessive, the question remains whether we would have been stumped due to the existence of multiple disease genes in the same region. HLA-loci and apolipoprotein genes are other examples of linked genes with similar function, where it has been difficult to identify which specific gene is involved in any given phenotype. Similar situations are likely to be frighteningly more common than we may expect, given that the location of genes in the human genome is certainly not random, as our mathematical models most often stipulate.

The most general model of allelic complexity is that of multiple disease-predisposing alleles in multiple unlinked genes (in Figure 2 this would look the same as Model 3 except the genes with disease-predisposing alleles would be on different chromosomes). This is probably the best supported general model of the majority of complex phenotypes based on what we know from lower organisms [13] and on general evolutionary principles [10,11]. Two examples of extreme locus and allelic heterogeneity for relatively simple phenotypes are nonsyndromic hearing loss [16•] and retinitis pigmentosa (see [15]). Truly common complex phenotypes will certainly involve multiple genetic and environmental risk factors. Whether or not association mapping will work in practice for these multifactorial phenotypes is highly dependent on both the allelic architecture of disease, the existence of detectable linkage disequilibrium, and the number of loci involved in disease predisposition, not to mention the environment, or interactions therewith.

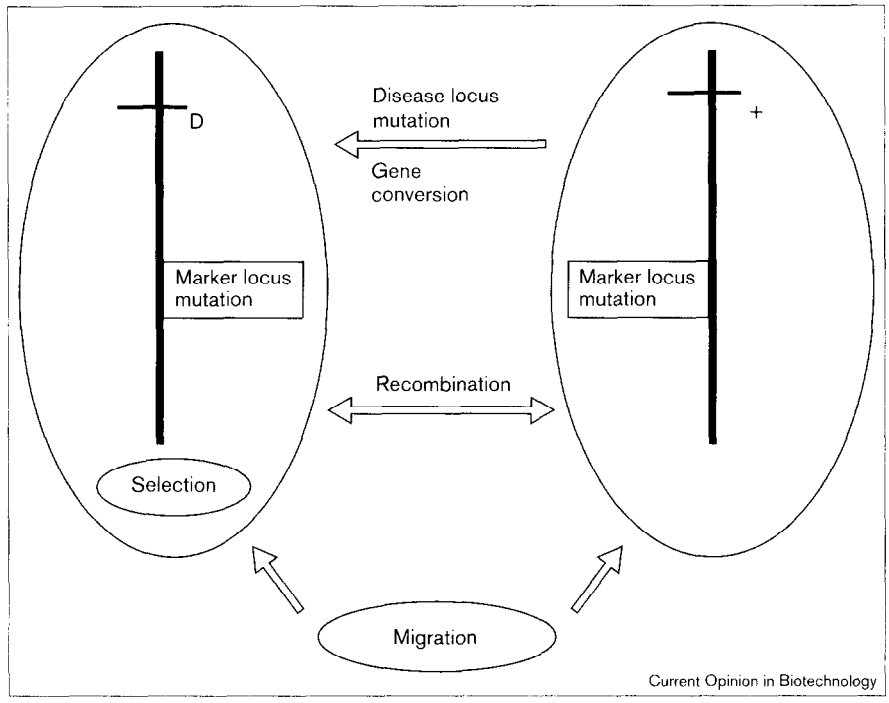
Where does linkage disequilibrium come from?

LD can be defined as the nonrandom assortment of alleles. Of particular interest are associations between sites tightly linked on the same chromosome. Such LD appears in the form of differences in the allele frequency distribution of one locus conditional on the alleles present on the same chromosome at another locus. If one locus has a disease-predisposing allele, and this allele is in LD with alleles of nearby marker loci, this phenomenon can be exploited to map the disease gene. The existence of LD does not mean there has to be a single haplotype associated with the disease-predisposing allele, nor does it imply necessarily that there has to be a single disease-predisposing allele, though for rare highly-penetrant monogenic diseases this is the most familiar presentation. Aspects of the population history, unrelated to any disease, however, play a critical role in the probable efficacy of LD mapping. For example, admixture can generate LD even among loci on different chromosomes, until sufficient generations pass for recombination to remove the association [17]. This kind of association makes less difference in pedigree data than in analysis by LD in populations. For a detailed summary of the forces that create and destroy LD, the reader is referred to [18••], the basics of which are summarized in Figure 3.

Under any model of allelic architecture, the forces of evolution act on chromosomes that carry D as one population of chromosomes, and the remaining larger set of chromosomes not carrying disease-predisposing alleles (+) evolve as a separate population in a small neighborhood around the disease locus. Recombination homogenizes these chromosomal populations through reciprocal gene flow, where a recombination exchanges chromosomal segments between these two groups. At large genetic distance from the disease-predisposing alleles, the chromosomal evolution is largely independent of the disease-predisposing alleles. Mutation that creates new disease-predisposing alleles acts as one-way gene flow taking a chromosome from the + population into

Figure 3

The forces that create and destroy linkage disequilibrium are shown graphically. Chromosomes with a disease-predisposing allele 'D' and chromosomes carrying a non-disease-predisposing allele '+' evolve as independent populations of chromosomes in a neighborhood around the disease locus. Recombination acts as reciprocal gene flow between these populations, homogenizing them; disease-locus mutation and gene conversion act to bring whole chromosomes from the '+' population into the 'D' population. Selection and marker-locus mutation act independently on 'D' and '+' chromosomes, increasing the diversity between them. Migration alters the composition of 'D' and '+' populations to a different degree, conditional on the frequency of 'D' and '+' in the migrants. For more details, see [18••].



the D population. Gene conversion has the same effect on LD as disease locus mutation, as it would take a D allele and place it on a chromosome from the + population. Marker locus mutation can increase the difference between the D and + populations because it occurs independently in the two groups. For microsatellite loci, the effects of marker locus mutation depend on whether or not the population of chromosomes in each group is sufficiently large for mutation-drift equilibrium (where marker allele frequency distribution is held constant as mutation creates new variation at the same rate with which drift eliminates existing variation) to stabilize the marker allele frequency distributions. For less mutable marker loci, mutation will create new alleles over time within each of these populations, and those new alleles will be the most striking markers of differentiation between the D and + chromosomal populations. Novel sequence polymorphisms arising as a result of recent mutation events will have similar properties. Migration and admixture can create LD as well. When two populations mixing together have different frequencies of D alleles, there will be different amounts of admixture in the D and + populations. This will lead to greater differences between the D and + populations in the admixed population, assuming the marker has different allele frequency distributions in the two mixing populations to begin with [17].

There is little debate about the points mentioned thus far. An issue that has led to some of the most violent arguments in population genetics has been the importance of selection. If an allele conferred a selective advantage at some point in history, that allele may have increased in frequency, dragging along neighboring neutral polymorphisms in a hitchhiking effect. If selection has played an important role in increasing the frequency of 'once rare' alleles that today predispose to some complex phenotype, then it is possible that there is increased allelic homogeneity. Examples where such a selective model has been hypothesized include the alleles leading to hemochromatosis, cystic fibrosis, Tay Sachs, and many others. It has also been argued that the data for those genes are consistent with the predictions of neutral theory (i.e. most genetic variation occurs without selection for or against it) [19••], as genetic drift might cause one of many rare D alleles to increase in frequency relative to the others. By contrast, the effect of positive selection might be to increase the frequency of all D alleles, leading to maintenance of high levels of allelic diversity. Interestingly, most of these diseases are much more frequent in one population (e.g. Europeans) than elsewhere, and it is not clear whether selective forces may also have been geographically restricted. In fact, in all of these cases, these is considerable allelic heterogeneity within and among populations, but it has not been possible to make convincing arguments about the relative likelihood of selection versus drift because too many parameters of human population history are not known.

How much LD is there?

The first question one must ask is how much LD exists in a given population and over what distances. In an attempt to address this issue, a number of investigations have been

done to examine the extent and nature of LD in different populations for different types of marker loci, for examples see [20•,21,22•,23••]. In each of these cases, as expected, pairs of tightly linked markers showed more evidence of linkage disequilibrium than unlinked pairs of markers. This is a necessary prerequisite for LD mapping of disease-predisposing loci to work. This is generally what is seen, but at closer distances the relationship between genetic and physical distance is no longer always simple or even monotonic, a potentially important issue in regard to fine-scale gene mapping.

Nickerson *et al.* [22•] and Clark *et al.* [23••] have examined a 9.7 kilobase (kb) region of the human lipoprotein lipase gene and found substantial amounts of polymorphism in both coding and noncoding sequence across a sample of 142 chromosomes from three populations, one of which was North Karelian Finns. In this study, the average individual was heterozygous for about 17 sites over this 9.7 kb region, with a total of 88 polymorphic sites found in the sample. Based on an analysis of the pairwise LD among the 2211 marker pairs studied in this sample, the authors concluded that "it is not the case that all such SNPs will give reliable information about flanking sites" [23••]. There was a substantial number of marker pairs that showed striking amounts of linkage disequilibrium, and at least three haplotypes were present in distantly related populations. Despite the fact that there was often LD between very closely spaced markers, the number of pairs for which no LD was detected was sufficiently large, and the cladistic network of haplotypes [24,25] was a sufficiently "uninterpretable tangle of loops" that the authors concluded that "a blind association or disequilibrium test with three or four random markers chosen from the variable sites within 10 kb of the lipoprotein lipase gene would not be a reliable way to detect nearby causal variation" [23••].

Some of the haplotype complexity Clark *et al.* [23••] observed was suggested to have possibly arisen as a result of earlier gene conversions. Other reports of gene conversion and theoretical analyses of gene conversion and its ramifications have been published recently [26–32]. Reports have been made of what appear to be recurrent mutations but might be explained as the result of an ancient gene conversion event [33]. Also, examples exist where the length of a conserved haplotype around a rare recessive disease allele in a young, isolated population was vastly shorter than what would be expected due to historical recombination alone — which might also be due to gene conversion (see, for example, [34]).

Although it is clear there is a greater chance to observe LD between alleles of tightly linked loci, it is unclear whether LD is sufficient to allow identification of the alleles predisposing to complex disease for many reasons: firstly, disease locus genotypes are difficult to infer because the genotype–phenotype correlations are far from deterministic and allelic heterogeneity is not unlikely to be abundant; secondly, man is a diploid organism, and it may not be easy to infer marker haplotypes from marker genotypes, especially when individuals are multiple site heterozygotes and there is an essentially open-ended number of haplotypes in the population because of historical recombination and mutation; thirdly, even if the disease locus genotypes are known, it is not easy to determine the phase of heterozygotes at this locus relative to the marker haplotypes — even when each are known without error; and finally, if the disease-marker LD pattern one finds is old (e.g. parts of the lipoprotein lipase gene [23••]), it might be possible to extrapolate across populations, or from a pilot study, to a larger 'replication' sample, but if it is not, the pattern of LD will be very different across populations (see [20•]).

How much allelic complexity is there?

In the past few years, there has been much debate about the amount of allelic complexity we expect to find in the genes that predispose to complex disease (see [2•,12•]). To assess the empirical evidence for allelic heterogeneity in the etiology of genetic disease, a literature review was performed, in which the results of mutation analyses published in the *American Journal of Human Genetics* volumes 60–62 (January 1997–June 1998) were considered systematically. In Table 1, a list is given of the genes for which multiple unique molecular alleles predisposing to some pathological phenotype were described. For the genes listed, the number of alleles described is typically in the tens or hundreds of alleles, where the exact number is correlated with the prevalence of the phenotype and how thoroughly it has been studied. Numerous examples, with details, are available in the Human Gene Mutation Database (http://www.uwcm.ac.uk/uwcm.mg/hgmd0.html).

It is granted that the majority of known disease genes have alleles predisposing to monogenic disease. The majority of those listed in Table 1, therefore, are alleles with severe phenotypic consequences — often recessive in nature. Heterozygotes for alleles predisposing to severe recessive disease are not selected against, explaining why substantial allelic heterogeneity may come to exist through mutation and genetic drift. These data have been ascertained through affected individuals and we have little if any systematic data on the relative complexity of variation at the same loci in the general (affected or not-yet-affected) population. Weiss [10,11,12•] has argued that the allelic complexity might be even greater for diseases of late age of onset, as negative selection is not acting strongly on phenotypes that typically afflict individuals later in life, after reproduction has taken place. At this point there is not much data about the allelic complexity of common pathologies in man, so it is possible that for some genes there may be a limited number of ways they can be mutated without leading to phenotypes that would be strongly selected against. If this is the case, there may be sufficiently few alleles in some genes for LD mapping to work, though for animal and human examples, the genes that are known do tend to show substantial heterogeneity [10,11,12•,13,35].

Kruglyak [36••] and Xiong and Guo [37] have similarly warned of the ramifications of allelic complexity on the hunt for disease-predisposing alleles by allelic association. The existing literature on searches for loci contributing to complex quantitative traits largely comes from agricultural and experimental genetics, but is not helpful on these points, as the experimental designs have not permitted an evaluation of intra-locus complexity in the general natural populations. Neither has the human mapping data for complex traits.

Beyond the issue of allelic heterogeneity is the mounting evidence that there is heterogeneity in the expression of different disease-predisposing alleles in the same genes. Molecular mechanisms involved in some splice-site alleles have been proposed by Rave-Harel *et al.* [38] for cystic fibrosis and Gantla *et al.* [39] for Crigler-Najjar syndrome type I. Differential expression for different deletions was shown for retinoblastoma [40], multiple alleles of the X-linked G4.5 gene were implicated as causal agents for different infantile dilated cardiomyopathies [41], and multiple investigations were conducted on the effects of different alleles in various genes in genotype–phenotype correlation studies (e.g. [25,35,42–49]). Methods are being developed to deal with sorting through this complexity, though inferring causality of a given allele is not a simple process even when the gene locus itself is known [50•].

Despite the mounting evidence for substantial heterogeneity in the disease-predisposing allelic spectrum, most statistical methods have been developed and evaluated under the implicit reductionist assumption that complex disease is caused by multiple genes, each having a single D and single + allele. Why is this the case, despite the mounting evidence of the 'biochemical individuality' of each person [12•]? Aside from the mathematical intractability of these models, this is largely because it is acknowledged that only if this assumption holds (or at least if there is minimal allelic heterogeneity) will LD be useful in detecting genes [37]. As stated by Kruglyak [36••], "there is little or no comparable [to the case of rare monogenic disease in isolated populations] evidence [that LD is detectable in any population] for common genetically complex disorders", and that "great care will be required if reality is to be distinguished from wishful thinking."

Population genetic methods for LD mapping in the presence of allelic heterogeneity

Laan and Pääbo [20•] have investigated empirically the earlier hypothesis [51,52] that in populations that have not undergone a recent demographic expansion, there may be greater levels of LD than in exponentially expanded populations. They examined several populations of varying size and structure and demonstrated that pairwise LD between microsatellite loci was much more striking in the Saami (non-expanded population) than in Finns (rapidly expanded population). In light of this observation, Laan and Pääbo [20•] proposed that the LD generated by drift in small populations of static size could be used to map genes for common disease ('drift mapping'). Freimer *et al.* [53] responded to this proposal by pointing out that one would not be able to detect shared segments or shared haplotypes around disease alleles in such populations, because there would be too much background LD in the population, clouding the interpretation. Terwilliger *et al.* [18•] subsequently summarized the theoretical background to such 'drift mapping' in extreme population isolates, demonstrating that whereas it is true that haplotypes are not expected in small, non-expanding populations, there will be higher levels of LD between disease alleles and linked marker alleles because over time genetic drift constantly creates new LD faster than the forces of recombination and mutation can make the LD decay. They proposed that even in the presence of substantial allelic heterogeneity, over time LD will be generated in populations of small effective population size (N_e) making it possible for LD mapping to work. This theory was then applied to the human renin-binding protein on the X-chromosome, which was successfully mapped in the Saami to the correct chromosomal region, whereas there was no detectable LD in a comparable Finnish sample [54••]. It was further pointed out [18••,54••] that the background LD between the marker loci actually decreases the false positive rate, as it increases the autocorrelation in the LD test statistic between linked marker loci, which has the expected effect on decreasing the effective number of independent tests [55]. Other ongoing studies have been reported using populations with similarly small N_e in the recent literature, for examples see [56,57]. The importance of these results is that in some situations allelic heterogeneity and ancient genesis of a disease-predisposing allele are not insurmountable problems, if the study makes full and appropriate use of population genetic theory.

Another population genetic approach — independent of allelic heterogeneity — that has been further developed in the past year [18••,58,59•,60] is the idea of admixture mapping [17]. When individuals from two genetically very different populations mate, the next generation will have substantial amounts of LD between both linked and unlinked pairs of loci. Over time, the LD will decay, but much more rapidly for unlinked locus pairs than for tightly linked ones. Optimal conditions for this method exist when there is minimal genetic variation within either parent population, and maximal genetic variation between them. For this reason, heterogeneous populations, such as African Americans, are not well suited to this approach, as there is so much genotypic and phenotypic variability within the parent populations, as well as in the admixed group [61]. Populations such as the Greenlanders, however, might be more ideal, as the parent populations (Inuit and Danes) are well defined, yet very different genetically, and the Inuit are among the more homogeneous of human populations [62–64]. Furthermore, the Greelandic population has not expanded very rapidly in recorded history, allowing drift mapping and admixture mapping to

combine forces. One of the most important features of admixture mapping, much as in drift mapping, is that allelic homogeneity is not an essential requirement for this technique to work, though the more homogeneous the parental populations are, the higher the power will be. Of course, to be useful for mapping a particular disease, that disease must be present in sufficient frequency in the admixed population; this may not always be the case in more extreme isolates.

One analytical approach to dealing with allelic heterogeneity within sequenced candidate genes would be to simply screen the entire coding sequence (and where known, the regulatory regions) of a gene in a large case-control sample and identify all variants. If there are disease-predisposing alleles in the gene, one would expect there to be an increased amount of nonsilent polymorphism in the affected individuals, when compared to controls. The background variation levels (on the + chromosomes) would be the same, and the affecteds would be enriched for the additional D alleles predisposing to the disease. For this reason, the overall number of detectable differences between affecteds and some consensus sequence should be larger than the overall number of differences between controls and the same consensus sequence (see [50•]). Such approaches would require the development of practical mass-sequencing technologies. More complicated models based on the type of sequence variation, its predicted effect on protein structure or regulation, and so on, can be modeled into such an analysis, once the nature of DNA sequence variation and its effect on protein structure and function becomes more accurately predictable. This is an important area of research that will be developed more in the coming years. Other purely analytical approaches have been proposed that allow for allelic heterogeneity by estimating the strength of LD along a chromosomal region using a variety of statistical analysis techniques (for example, see [37,65,66,67•]).

Homogeneous populations are not a panacea

The diseases of the Finnish disease heritage are classical examples where LD mapping has worked fantastically, in that haplotype analysis and shared segment analysis has led to the cloning of many rare recessive disease alleles. Because of the small size of the founder population and its subsequent rapid expansion, with a paucity of immigrants due to cultural and geographical isolation, a number of rare recessive disease-predisposing alleles have increased in frequency to the point where homozygotes are not infinitesimally rare. Even though the diseases caused by these alleles are more common in Finland than in other parts of the world [5], they retain sufficient allelic homogeneity for haplotype mapping to work.

In trying to generalize this result to more common complex disease, numerous problems have been encountered. Not the least of them is that the Finnish population is not dramatically less heterogeneous than other populations [22•]. Also very important is that if the prevalence of a common disease is 10%, the predisposing alleles (which do not cause disease, but increase by a small factor the likelihood of getting a disease) must be much more frequent than this. In this case, there would be multiple lineages for the disease-predisposing allele(s) of any locus, even in a population as young and homogeneous as Finland. If the frequency of the disease-predisposing alleles combined were 0.1 and there were 5000 founders, in expectation there would have been 1000 D alleles at the time the population was founded, and thus multiple lineages within the recorded population history are expected. The rapid population expansion will further have prevented the generation of significant amounts of new LD over time [18••]. This contrast between common and very rare disorders is exactly what is seen in other well studied populations, such as French Canadians [7,8].

Even in Finland, much more extreme local population isolates have been studied in the search for complex disease genes by allelic association methods. An example of this is the recent ongoing study of schizophrenia in the isolated Northeastern community of Kuusamo [68,69•]. In this region, there were 80 founders about 10–15 generations ago, rapidly expanding to a current population of 18,000, with negligible amounts of immigration [68]. In this study, several important problems related to the study of allelic association in extreme population isolates became obvious (see [69•,70]). Perhaps their most generally relevant observation was that in any study of a disease which clusters in families, in a finite closed population, any sample of affected individuals must be more closely related to each other than to any control sample of unaffected individuals. If one thinks of LD as genetic differentiation (measured by F_{st}, for example) between a population of affecteds and a population of controls (see [54••,70]), then the null hypothesis in such a case-control study would not be that $F_{st} = 0$, but would rather be that $F_{st} = c$, where c is some value greater than 0, due to the incumbent greater genetic similarity and relatedness of cases to each other, even for marker loci unlinked to the disease loci. This is exactly what was found in the genome scan undertaken by Hovatta *et al.* [69•]. It is important to take this possibility into account in interpreting the results of association analyses in small population isolates, as there may be unavoidable tendencies toward false positives. To this end, when conducting a genome scan, the interpretation of the statistical findings are best made cautiously.

Population genetic epidemiology

Another approach that is gaining momentum is the idea of cross-population studies, in which association studies are done in a comparative manner across multiple populations simultaneously. This is related to what was done in Finland, where the isolate studied for schizophrenia was selected based on a geographical epidemiological analysis of the prevalence in different regions of the country [68]. Phenotypic variation between populations should be

correlated with genetic variance between populations, if a disorder has any substantial genetic risk alleles. Human populations have a defined, if not yet well-characterized, evolutionary interrelatedness, and this can be used as a tool to decipher the complexities of human phenotypic variation. Systematic studies of these genetic differences between human populations are ongoing [71], and a human genome diversity project has been proposed to coordinate these efforts (see [12•,72••]). The SNP map currently under development [2•] will be of much greater utility in tracing the history of human populations [12•,73] than it is likely to be for the dissection of complex traits through association mapping, but this information can ultimately be of great use to genetic epidemiologists. If a risk factor is identified for a given disease, then one could analyze its contribution to the overall phenotypic variance across multiple populations, comparing its effect in different populations with the other sources of genetic and environmental variation between those populations. This information can be useful for targeting which populations should be studied more extensively to search for other remaining etiological factors of importance, for which the studied risk allele does not explain much of the trait heritability. This kind of analysis, however, is difficult even for well understood genes, such as ApoE [74], and is highly vulnerable to the well known 'ecological fallacy' of epidemiology, that is, many correlated but unmeasured risk factors may also vary among populations.

In the past year, several methodological and applied papers have begun to propose the simultaneous use of multiple populations in a controlled manner to separate the wheat from the chaff. Merriman *et al.* [75] suggested that investigators should collect "large numbers of families from multiple populations that should be as genetically homogeneous as possible" in the effort to identify complex disease-predisposing alleles. McKeigue [58,59•] has proposed mapping disease-predisposing loci by estimating the ethnic background of different chromosomal regions in recently admixed populations. To apply this method in practice, one would need to have good genetic and phenotypic data about the parental homogeneous populations that mixed together, and about the admixed population. This proposal is a step towards population genetic epidemiology as well, in that the genetic differences between populations are compared with the phenotypic differences in a directed way to map disease-predisposing genes. Valdes *et al.* [76] performed an LD analysis of the HLA gene region and insulin-dependent diabetes mellitus in another set of human populations simultaneously to look for differential effects in different genetic and environmental backgrounds. Again, the use of multiple populations jointly appears to be a more general movement in the field towards population genetic epidemiology, examining genotype and phenotype correlations jointly through a combination of geographic epidemiology and population genetics.

The trends are starting to move in this direction, and the importance of having good data about the genetic interrelationships between populations and good epidemiological data across populations is becoming more apparent. If the SNP map currently being developed, at the cost of hundreds of millions of dollars of taxpayers money, is to be of real benefit in the future understanding of complex human phenotypes, it may well be through its impact on the ability to accurately quantify and describe the genetic differences between populations, and not for direct association mapping, which as Kruglyak [36••] points out, may well be mostly wishful thinking.

One statement that is often taken grossly out of context by well meaning medical geneticists is the statement of Cavalli-Sforza and Cavalli-Sforza [77] that "[aside from the external differences between human populations] the remainder of our genetic makeup hardly differs at all." A varient of this is that 85–90% of all human genetic variation is found within, not between, ethnically defined groups. They are not saying that the genetic difference between populations are not critical to account for in gene-mapping studies, but are rather pointing out that the sequence diversity is extremely low per nucleotide across the entire species, in an effort to make a political statement. Some medical geneticists have taken this statement out of context to mean that because there is more variation within populations than between populations, they do not need to worry about the homogeneity of their study sample in an association study. This kind of analysis, however, is difficult even for well understood genes, such as ApoE [74], and is highly vunerable to the well known 'ecological fallacy' of epidemiology, that is, many correlated but unmeasured risk factors may also vary among populations. In an objective scientific analysis, it is clear that if we are searching for genetic factors that contribute to the variance of any trait, we want to collect our study samples in a way that minimizes the overall genetic variance not related to the trait. Thus, if there is any genetic variance between populations, the power of a study would increase if this variance were eliminated by concentrating on a homogeneous group. The less non-trait-related genetic variance the better, and clearly there are enough differences between populations to justify research to quantify and document this variation (see [71]). Gene mappers are painfully aware of the great inter-population variation related to the allelic architecture of complex traits. If a phenotype has the same prevalence across populations, there is typically not going to be much power in looking for major genetic factors, as there are not very many genetic (or environmental) factors of constant frequency across populations. Again, whereas alleles in the same genes or genes in the same biochemical pathway may be involved across populations, the specific alleles involved in different populations may be substantially different — as is the case for the alleles of the BRCA1 and BRCA2 loci, among others [78].

Joint testing of linkage and LD

There are a few issues that need to be addressed regarding the statistical analysis of LD. There is a popular belief that the transmission/disequilibrium test (TDT) [79] is an LD test and that a positive result from such a test rejects the null hypothesis that there is no allelic association. This is only the case, however, when the entire study sample consists of singleton (i.e. unrelated) affecteds and their parents, a point that has been made numerous times in the literature (see [80–82]). Recently there have been numerous modifications to the TDT design, including using other relatives as controls when parents are dead [81–86], looking at multiple allelic markers [81,87], or extending the method to handle quantitative trait loci [88,89•]. Most of these extensions start by considering the simple case where there is one affected individual per pedigree, and all pedigrees are mutually unrelated. In this scenario, the TDT is equivalent to the haplotype relative risk McNemar test for LD [80], and so a positive test would reject the null hypothesis of no allelic association. As soon as there is more than one affected individual per pedigree, however, the only null hypothesis a significant TDT can reject is that of no linkage [80–82]. Although it is true that the power of the TDT is larger when there is LD, in a set of sib-pairs the TDT will have power to detect linkage in the absence of LD (JD Terwilliger and HHH Göring, unpublished data). In larger pedigrees, the power can be substantially greater as a linkage test in the absence of LD.

There have been numerous papers in the past year about joint analyses of linkage and LD that examine the phenomena jointly in a more powerful and statistically responsible manner [90•,91,92,93•,94•] (JD Terwilliger and HHH Göring, unpublished data)]. The essence of these can be summed up as follows. In simple terms, the TDT, haplotype-relative risk (HRR), and affected sib-pair (ASP) tests are each based on the common assumption that each parent is informative for the disease-predisposing allele (D/+ genotype), and that all affected individuals receive a D allele from each of their parents. In the TDT [79], as it is a linkage test, if the marker locus is homozygous, there is no information coming from that parent. In the haplotype-based haplotype-relative risk (HHRR) [80], a test of LD, the child's genotype is used to infer the phase in the parents to estimate the parental haplotype frequencies — or in other words, to test for the presence of allelic association between the trait and the marker in the parental generation. In the affected sib-pair method, transmission from heterozygous parents to affected offspring is analyzed, such that if there is linkage the affected children should share more marker alleles identical by descent than they would by chance. In those heterozygous parents, the possible disease-marker phases are assumed to be equally probable. By analogy, in a case control association study, contrasting allele frequencies between affecteds and controls, we are estimating the allele frequencies conditional on the affection status, implicitly assuming the cases and controls have different disease locus genotypes (e.g. D/D for affecteds and +/+ for controls). Further extrapolation of this model is provided by Terwilliger and Göring (unpublished data), who discuss each of the common statistics from this unified framework.

Interpretation of LD analyses

There are numerous different approaches to significance testing of LD, ranging from simple contingency table chi-square tests through to complex likelihood-based procedures. If strong enough LD exists, any of the methods should give similar results. A more important issue than how to do the analyses is how to interpret the results. Kruglyak [36••], Camp [95], Elston [96••], Morton [97••], Vieland and Hodge [98••] and others have written on the subject of how to interpret significance, each based on different theoretical and philosophical models. There are as many such means of interpretation as there are statisticians. Personally, we prefer to think of the situation by analogy to gambling — in that ultimately it is the investigator doing the research who will have to decide at what point a finding is sufficiently significant that he wishes to put his own time and money at risk to follow it up with the expensive and time-consuming labor that comes after the detection of LD or linkage. Some people like to gamble in the hope that they might win the jackpot, but so long as they realize what the odds are that they will lose their shirt, all of that is fine. For that reason we reject the idea of uniform codes or rules that everyone must stick to in the interpretation of their own findings (e.g. [55]). A very impotant and unfortunate point about the socialogy of science, especially in competitive funding times, however, is that review panels tend to be conventional sometimes to the point of ritual, insisting on fixed criteria (power calculations, multiple test corrections, etc.) for awarding funds.

A cursory examination of the literature makes one quickly become jaded about the way people interpret their results, and it makes one think that people do not sufficiently appreciate how heavily the odds are stacked against them in LD mapping. It is widely believed that there is a publication bias (see [99]), in that one might tend to report something in the literature primarily in two situations — either when obtaining a significant finding, or when trying to replicate a positive finding someone else presented. One would thus expect a systematic tendency toward significant p-values being reported more often than expected by chance, even if none of them were real. This is because most negative results are not expected to be published at all, as they are neither surprising nor interesting to most readers. In order to examine this and give the reader a means by which to judge his own interpretation of significance, I went through every article published in 1997 in either the *American Journal of Medical Genetics* (*Neuropsychiatric Genetics*) or *Psychiatric Genetics*, two journals that publish a large number of association studies. Over the past two years there were a total of 222 p-values reported, and there were an additional 39 tests reported simply as 'nonsignificant at the 0.05 level'. A histogram of those

p-values (mostly from candidate gene studies, attempts to replicate earlier findings, or newly reported 'significant' associations) is shown in Figure 4. The distribution looks almost uniform, in contrast to what might have been expected if there were a publication bias. A goodness of fit test of this data to a uniform distribution showed a very good fit ($p > 0.87$), leading to the possible interpretation that even the 'significant' reports may not be real, as there are just as many p-values that are too large than those that are 'too significant.' A p-value of 0.01 may look significant, but in the context of the multitude of tests being performed, and the independence of each test under the null hypothesis, two main things are indicated: firstly, investigators are too frequently gambling on and publishing results in situations where the evidence is not at all compelling; secondly, the recent advocacy for association analysis as the savior of complex disease gene mapping (e.g. [4]) is leading investigators to invest prematurely in this strategy, which has thus far led to a whole lot of nothing — certainly in these psychiatric genetics studies surveyed.

Conclusion

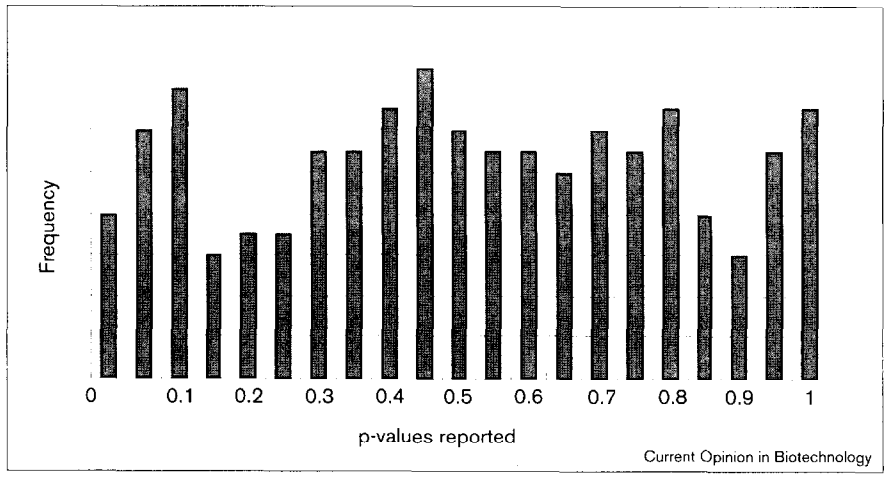
All of this work begs the question of what is the most powerful approach to detecting disease-predisposing loci. Linkage analysis remains a favored method of choice whenever large pedigrees are available (for examples, see [18••,37,96••,97••,100]), as it is not dependent on allelic homogeneity and does not require an infinitely dense map of marker loci. In fact, for linkage analysis, the overlyhyped SNP map [2•,36••,101,102] is probably not a practical tool, for numerous reasons beyond the scope of this review (see [70]). Furthermore, as one increases the sample size, the power of a linkage approach increases, whereas for an association analysis this is not necessarily the case. Testing up to 100,000 SNP markers in a given study, for example, raises serious questions about how to judge the thousands of results that will appear to be positive by standard p-value criteria (or will have relatively high likelihood ratios for those who prefer them [99]). Can enough cases and controls even be identified to achieve convincing results? If there is substantial allelic heterogeneity, then as one increases the sample size, the number of different disease-predisposing alleles (each with their own independent haplotype of nearby marker alleles) may likewise increase, and thus there may never be much power even with complete ascertainment of the entire human population. Obviously this is not a desirable property for any statistical analysis method.

To conclude this review, we want to state clearly that we are not contending that association analysis does not have its place. It is one tool in the arsenal of geneticists who are engaged in attacking a very tough problem — unraveling the puzzles that millennia of evolution have assembled for us in the form of complex phenotypes. Rather than focusing on statistical methods to improve the power of detection, which are reviewed elsewhere (see [103••,104•]). I have focused on an analysis of the more fundamental question — when will allelic association exist, and how can we use population genetics to identify those situations where LD mapping may be an appropriate weapon of choice. Whereas penicillin has been an important weapon in the war on certain types of bacterial infection, it is not so useful in fighting breast cancer. Similarly, whereas allelic association has been a great tool in unraveling the secrets of rare recessive diseases in isolated populations, there is no empirical evidence that it will be as useful in decoding the genotype–phenotype relationships in complex human traits, and certainly not in cosmopolitan heterogeneous populations.

Our hunch is that too many people are concentrating on simple mathematically tractable models that assume the only difference between simple disease and complex disease is related to effect size of a single allele per locus, whereas there is a looming danger that there is also a substantial increase in complexity in both allelic and non-allelic heterogeneity, gene by environment interactions, epistasis,

Figure 4

The distribution of all reported p-values from association studies in either *American Journal of Medical Genetics* (*Neuropsychiatric Genetics*) or *Psychiatric Genetics* in 1997 is shown. A total of 222 reported p-values are graphed in the figure, and an additional 39 tests were listed as 'nonsignificant' at the 0.05 level with no statistical details in the manuscript. If all of the results were obtained under the null hypothesis, the expected distribution would be uniform. As can be seen in this figure, there is very good fit to the uniform expectation ($\chi^2_{(20)} = 12.98$; $p > 0.87$), indicating that the published p-values are consistent with the absence of gene effects in all the published analyses.



pleiotropy, and variable expressivity of different alleles in the same gene. In the end, it is important to re-evaluate on an individual basis what you believe about the genetic basis of the phenotype you are studying — starting from first principles, so that you can see exactly what are the critical assumptions that form the basis of any disagreements you may have with others. It is ultimately important that we do not look to the coming SNP map as a panacea, but rather that we retain a sense of cautious optimism at best. Let us conclude with a return to the sage advice of Confucius who advised, "Do not be impatient, do not look for small gains. Wish for haste and you will not accomplish your objectives. Look for small profits and the important tasks will not be accomplished." (Lun Yu 13:17) [1].

Acknowledgements

A Hitchings-Elion fellowship from the Burroughs-Wellcome Foundation is thankfully acknowledged. Thanks to Matthew R Clark for critical reading of this manuscript, and helpful comments on the ideas, structure, and organization of this review. Harald HH Göring is likewise acknowledged for his helpful comments on an earlier version of this manuscript and for hours of thoughtful discussion of the philosophical and scientific issues presented in this review.

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. Confucius: *Lun Yu (Analects). Simplified Chinese Edition.* Hong Kong: Confucius Publishing Company; 1996.

2. • Chakravarti A: **It's raining SNPs, hallelujah.** *Nat Genet* 1998, **19**:216-217.
This 'News and Views' article presents Chakravarti's somewhat more optimistic viewpoint about the utility of an SNP map for identification of alleles predisposing to human complex disease.

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

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

5. Nevanlinna HR: **The Finnish population structure: a genetic and genealogical study.** *Hereditas* 1972, **71**:195-236.

6. Puffenberger EG, Hosoda K, Washington SS, Nakao K, deWit D, Yanagisawa M, Chakravat A: **A missense mutation of the endothelin-B receptor gene in multigenic Hirschspung's disease.** *Cell* 1994, **79**:1257-1266.

7. Heyer E, Tremblay M: **Variability of the genetic contribution of Quebec population founders associated to some deleterious genes.** *Am J Hum Genet* 1995, **56**:970-978.

8. Heyer E: **Genetic consequences of differential demographic behaviour in the Saguenay region, Québec.** *Am J Phys Anthropol* 1995, **98**:1-11.

9. Kaplan NL, Hill WG, Weir BS: **Likelihood methods for locating disease genes in nonequilibrium populations.** *Am J Hum Genet* 1995, **56**:18-32.

10. Weiss KM: *Genetic Variation and Human Disease.* Cambridge: Cambridge University Press; 1995.

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

12. • Weiss KM: **Perspective: in search of human variation.** *Genome Res* 1998, in press.
In this report, Weiss compares the ways in which the SNP map being generated will benefit the gene mapping community and the anthropological genetics community. His opinions about the complexity of phenotypic variation and the importance of studying interpopulation genetic variation are clearly stated, together with his cautions about the political concerns and issues that are of paramount importance to the Human Genome Diversity Project, and future studies of human population genetic epidemiology.

13. Mackay TFC: **The nature of quantitative genetic variation revisited: lessons from *Drosophila* bristles.** *BioEssays* 1996, **18**:113-121.

14. Bale AE: **Variable expressivity of patched mutations in flies and humans.** *Am J Hum Genet* 1997, **60**:10-12.

15. Ott J, Bhattacharya S, Chen JD, Denton MJ, Donald J, Dubay C, Farrar GJ, Felix JS, Fischman GA, Frey G *et al.*: **Localizing multiple X-linked retinitis pigmentosa loci using extended multi-locus homogeneity tests.** *Proc Natl Acad Sci USA* 1990, **87**:701-704.

16. • van Camp G, Willems PJ, Smith RJH: **Nonsyndromic hearing impairment: unparalleled heterogeneity.** *Am J Hum Genet* 1997, **60**:758-764.
One example is discussed here where a relatively simple phenotype has a whole ton of heterogeneity, with many different mutations identified in many different genes – each causing a similar phenotype. This is very similar to the situation of retinitis pigmentosa. One can only wonder if the genetics of other complex phenotypes will involve as many different loci as those for these vision and hearing related phenotypes.

17. Chakraborty R, Weiss KM: **Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci.** *Proc Natl Acad Sci USA* 1988, **85**:9119-9123.

18. •• Terwilliger JD, Zollner S, Laan M, Pääbo 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.
This manuscript provides an intuitive review of the forces that create and destroy LD by analogy to haploid population genetics theory. Further simulations and theoretical analyses are presented to demonstrate the utility of LD mapping in small populations which have not expanded, and in which there will be substantial LD in the absence of shared haplotypes.

19. •• Thompson EA, Neel JV: **Allelic disequilibrium and allele frequency distribution as a function of social and demographic history.** *Am J Hum Genet* 1997, **60**:197-204.
This paper presents an argument that the observed haplotype sharing data from diseases such as cystic fibrosis and hemochromatosis are consistent with neutral evolution. Furthermore, association mapping and population genetics are gradually merging back together, where they came from originally, and this paper provides a nice accessible interface between the active areas of research in both fields.

20. • Laan M, Pääbo S: **Demographic history and linkage disequilibrium in human populations.** *Nat Genet* 1997, **17**:435-438.
Empirical evidence is presented in this manuscript that there is substantially more LD in the Saami, a population which has been of stable size for many generations, than there is in the Finns, Swedes, or Estonians – populations which have undergone a rapid population expansion in recent history. This paper was the first to propose the idea that the LD generated by genetic drift could be used for gene mapping.

21. Perez B, Desviat LR, Ugarte M: **Analysis of the phenylalanine hydroxylase gene in the Spanish population: mutation profile and analysis with intragenic polymorphic markers.** *Am J Hum Genet* 1997, **60**:95-102.

22. • Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF: **DNA sequence diversity in 19.7 kb region of the human lipoprotein lipase gene.** *Nat Genet* 1998, **19**:233-240.
This manuscript describes the amount of polymorphism and LD between polymorphisms within a small section of the lipoprotein lipase gene. The levels of LD are surprisingly low, and the levels of polymorphism, especially in coding sequence, may be higher than expected. It is an important paper to read before embarking on any large scale association screens, as it is one of the best documented studies in this very important embryonic area of research.

23. •• Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: **Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, in press.
This paper presents an analysis of the data from [22•], focusing on the amount of LD exhibited between pairs of polymorphic loci in part of the human lipoprotein lipase gene. The authors also applied techniques of cladistic analysis to analyze the haplotype structure of the observed sequences across this small genomic region, and found an 'uninterpretable tangle of loops.' They went on to discuss the potential ramifications of these observations on association studies for complex human phenotypes.

24. Sing CF, Havilland MB, Zerba KE, Templeton AR: **Application of cladistics to the analysis of genotype-phenotype relationships.** *Eur J Epidemiol* 1992, **8(suppl 1)**:3-9.

25. Templeton AR: **Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome.** *Ciba Found Symp* 1996, **197**:259-283.

26. Betran E, Rozas J, Navarro A, Barbadilla A: **The estimation of the number and length distribution of gene conversion tracts from population DNA sequence data.** *Genetics* 1997, **146**:89-99.

27. Burghes AHM: **When is a deletion not a deletion? When is it converted?** *Am J Hum Genet* 1997, **61**:9-15.

28. Campbell L, Potter A, Ignatius J, Dubowitz V, Davies K: **Genomic variation and gene conversion in spinal muscular atrophy: implications for disease process and clinical phenotype.** *Am J Hum Genet* 1997, **61**:40-50.

29. Giordano M, Marchetti C, Chiorboli E, Bona G, Momigliano Richiardi P: **Evidence for gene conversion in the generation of extensive polymorphism in the promoter of the growth hormone gene.** *Hum Genet* 1997, **100**:249-255.

30. Hanneman WH, Schimenti KJ, Schimenti JC: **Molecular analysis of gene conversion in spermatids from transgenic mice.** *Gene* 1997, **200**:185-192.

31. Ohta T: **Role of gene conversion in generating polymorphisms at major histocompatibility complex loci.** *Hereditas* 1997, **127**:97-103.

32. Elliott B, Richardson C, Winderbaum J, Nickoloff JA, Jasin M: **Gene conversion tracts from double-strand break repair in mammalian cells.** *Mol Cell Biol* 1998, **18**:93-101.

33. Laiho E, Ignatius J, Mikkola H, Yee VC, Teller DC, Niemi K-M, Saarialho-Kere U, Kere J, Palotie A: **Transglutaminase 1 mutations in autosomal recessive congenital ichthyosis: private and recurrent mutations in an isolated population.** *Am J Hum Genet* 1997, **61**:529-538.

34. Pekkarinen P, Hovatta I, Hakola P, Jarvi O, Kestila M, Lenkkeri U, Adolfsson R, Holmgren G, Nylander PO, Tranebjaerg L *et al.*: **Assignment of the locus for PLO-SL, a frontal-lobe dementia with bone cysts to 19q13.** *Am J Hum Genet* 1998, **62**:362-372.

35. Sing CF, Havilland MB, Reilly SL: **Genetic architecture of common multifactorial diseases.** *Ciba Found Symp* 1996, **197**:211-232.

36. •• Kruglyak L: **What is significant in whole-genome linkage disequilibrium studies?** *Am J Hum Genet* 1997, **61**:810-812.
In this paper, some major issues of interpretation of genome scans for LD are discussed and elucidated in the context of the conservative tradition of the often cited Lander and Kruglyak [55] paper regarding linkage analysis. This paper is important reading as it summarizes some of the difficulties and obstacles to gene mapping through the use of LD.

37. Xiong M, Guo S-W: **Fine-scale genetic mapping based on linkage disequilibrium: theory and applications.** *Am J Hum Genet* 1997, **60**:1513-1531.

38. Rave-Harel N, Kerem E, Nissim-Rafinia M, Madjar I, Goshen R, Augarten A, Rahat A, Hurwitz A, Darvasi A, Kerem B: **The molecular basis of partial penetrance of splicing mutations in cystic fibrosis.** *Am J Hum Genet* 1997, **60**:87-94.

39. Gantla S, Bakker CTM, Deocharan B, Thummala NR, Zweiner J, Sinaasappel M, Chowdhury JR, Bosma PJ, Chowdhury NR: **Splice-site mutations: a novel genetic mechanism of Crigler-Najjar syndrome type 1.** *Am J Hum Genet* 1998, **62**:585-592.

40. Bremner R, Du DC, Connolly-Wilson MJ, Bridge P, Farid-Ahmad K, Mostachfi H, Rushlow D, Dunn JM, Gallie BL: **Deletion of RB exons 24 and 25 causes low penetrance retinoblastoma.** *Am J Hum Genet* 1997, **61**:556-570.

41. D'Adamo P, Fassone L, Gedeon A, Janssen EAM, Bione S, Bolhuis PA, Barth PG, Wilson M, Haan E, Orstavik KH: **The X-linked gene G4.5 is responsible for different infantile dilated cardiomyopathies.** *Am J Hum Genet* 1997, **61**:862-867.

42. Browne CE, Dennis NR, Maher E, Long FL, Nicholson C, Sillibourne J, Barber JCK: **Inherited interstitial duplications of proximal 15q: genotype-phenotype correlations.** *Am J Hum Genet* 1997, **61**:1342-1352.

43. Johnston J, Kelley RI, Feigenbaum A, Cox GF, Iyer GS, Funanage VL, Proujansky R: **Mutation characterization and genotype-phenotype correlation in Barth syndrome.** *Am J Hum Genet* 1997, **61**:1053-1058.

44. Kayaalp E, Treacy E, Waters PJ, Byck S, Nowacki P, Scriver CR: **Human phenylalanine hydroxylase mutations and hyperphenylalaninemia phenotypes: a metanalysis of genotype-phenotype correlations.** *Am J Hum Genet* 1997, **61**:1309-1317.

45. Cavalier L, Ouahchi K, Kayden HJ, di Donato S, Reutenauer L, Mandel J-L, Koenig M: **Ataxia with isolated vitamin E deficiency: heterogeneity of mutations and phenotypic variability in a large number of families.** *Am J Hum Genet* 1998, **62**:301-310.

46. Gilad S, Chessa L, Khosravi R, Russell P, Galanty Y, Piane M, Gatti RA, Jorgensen TJ, Shiloh Y, Bar-Shira A: **Genotype-phenotype relationships in Ataxia-telangiectasia and variants.** *Am J Hum Genet* 1998, **62**:551-561.

47. Jean-Pierre C, Denamur E, Henry I, Cabanis MO, Luce S, Cecille A, Elion J, Peuchmaur M, Loirat C, Niaudet P *et al.*: **Identification of constitutional WT1 mutations in patients with isolated diffuse mesangial sclerosis, and analysis of genotype/phenotype correlations by use of a computerized mutation database.** *Am J Hum Genet* 1998, **62**:824-833.

48. Manning BM, Quane KA, Ording H, Urwyler A, Tegazzin V, Lehane M, O'Halloran J, Hartung E, Giblin LM, Lynch PJ *et al.*: **Identification of novel mutations in the Ryanodine-receptor gene (RYR1) in malignant hypothermia: genotype-phenotype correlation.** *Am J Hum Genet* 1998, **62**:599-609.

49. Zhao HG, Aronovich EL, Whitley CB: **Genotype-phenotype correspondence in Sanfilippo syndrome type B.** *Am J Hum Genet* 1998, **62**:53-63.

50. • Petersen GM, Parmigiani G, Thomas D: **Missense mutations in disease genes: a Bayesian approach to evaluate causality.** *Am J Hum Genet* 1998, **62**:1516-1524.
This is an initial approach for distinguishing phenotypically relevant mutations in disease genes. In the future, there will be more work needed in this area to help distinguish background polymorphism from that which is functionally relevant, and beyond this, to using predictive models of the effects of the alterations on 3-dimensional protein structure and function to further help distinguish the functionally relevant alleles from functionally neutral polymorphisms. This paper is a starting point for thinking about these models in future methodological advances.

51. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theor Appl Genet* 1968, **38**:226-231.

52. Slatkin M: **Linkage disequilibrium in growing and stable populations.** *Genetics* 1994, **137**:331-336.

53. Freimer NB, Service SK, Slatkin M: **Expanding on population studies.** *Nat Genet* 1997, **17**:371-373.

54. •• Laan M, Pääbo S: **Mapping genes by drift-generated linkage disequilibrium.** *Am J Hum Genet* 1998, in press.
This manuscript provided the first application of the 'drift mappng' [18••] technique to a real phenotypically active polymorphism in the renin binding protein genes. In this study, they were able to detect LD with nearby microsatellites in Saami but not in Finns, confirming the predictions of the earlier manuscript.

55. Lander ES, Kruglyak L: **Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.** *Nat Genet* 1995, **11**:241-247.

56. Palmer LJ, Pare PD, Faux JA, Moffatt MF, Daniels SE, LeSouef PN, Bremner PR, Mockford E, Gracey M, Spargo R *et al.*: **$Fc_\varepsilon R1$-β polymorphism and total serum IgE levels in endemically parasitized Australian aborigines.** *Am J Hum Genet* 1997, **61**:182-188.

57. Li L, Drayna D, Hu D, Hayward A, Gahagan S, Pabst H, Cowan MJ: **The gene for severe combined immunodeficiency disease in Athabaskan-speaking native americans is located on chromosome 10p.** *Am J Hum Genet* 1998, **62**:136-144.

58. McKeigue PM: **Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations.** *Am J Hum Genet 1997*, **60**:188-196.

59. • McKeigue PM: **Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture.** *Am J Hum Genet* 1998, **63**:241-251.
In this paper, McKeigue expands on some of the issues involved in admixture mapping. While this technique may be of limited utility in the real world as it presently stands, it may become more relevant and useful in the future, when

we have more available data about the genetic and phenotypic differences between populations. On the plus side, it is one of the few approaches that is not theoretically dependent on allelic homogeneity in the disease locus.

60. Kaplan NL, Martin ER, Morris RW, Weir BS: **Marker selection for the transmission/disequilibrium test, in recently admixed populations.** *Am J Hum Genet* 1998, **62**:703-712.

61. Chakraborty R, Kamboh M, Nwankwo M, Ferrell R: **Caucasian genes in American Blacks: new data.** *Am J Hum Genet* 1992, **50**:145-155.

62. Harvald B: **Breakup of an isolate.** *Arctic Med Res* 1988, **47**:41-42.

63. Harvald B: **The genetic epidemiology of Greenland.** *Arctic Med Res* 1989, **48**:171-174.

64. Edwards AWF: **The structure of the polar Eskimo genealogy.** *Hum Hered* 1992, **42**:242-252.

65. Terwilliger JD: **A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci.** *Am J Hum Genet* 1995, **56**:777-787.

66. Devlin B, Risch N: **A comparison of linkage disequilibrium methods for fine mapping.** *Genomics* 1995, **29**:311-322.

67. • Lazzeroni L: **Linkage disequilibrium and gene mapping: an empirical least-squares approach.** *Am J Hum Genet* 1998, **62**:159-170.
This paper presents a multilocus approach to LD mapping based on identity by descent inference from multiple marker loci jointly. The author theoretically divides chromosomes into those with any 'D' allele and those without a 'D' allele, thus in theory allowing for allelic heterogeneity, though as with any LD method, the power is very adversely affected when the allelic complexity is great.

68. Hovatta I, Terwilliger JD, Lichtermann D, Mäkikyrö T, Suvisaari J, Peltonen L, Lönnqvist J: **Schizophrenia in the genetic isolate of Finland.** *Am J Med Genet* 1997, **74**:353-360.

69. • Hovatta I, Varilo T, Suvisaari J, Terwilliger JD, Ollikainen V, Arajärvi R, Juovinen H, Kokko Sahin ML, Väisänen L *et al.*: **A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation suggesting multple susceptibility loci.** *Am J Hum Genet* 1998, in press.
This study highlights some of the ups and downs of LD mapping in extreme population isolates. In this case, several issues of more general concern are pointed out, and potential solutions are proposed. This is useful reading for anyone working in extremely small expanded populations.

70. Terwilliger JD: **Mapping genes predisposing to complex traits in extreme population isolates.** *CSC News* 1997, **February**:23-26.

71. Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes.* Princeton: Princeton University Press; 1993.

72. •• Harding RM, Sajantila A: **Human genome diversity – a project?** *Nat Genet* 1998, **18**:307-308.
The authors comment on the feasibility, viability, and political considerations involved in the proposed human genome diversity project. Clearly, investigations of human population history and variation are critical to the future of enetic epidemiology and gene identification, and this review covers some subtle important considerations in very eloquent style.

73. Mountain JL, Cavalli-Sforza LL: **Multilocus genotypes, a tree of individuals, and human evolutionary history.** *Am J Hum Genet* 1997, **61**:705-718.

74. Stengard J, Weiss K, Sing C: **An ecological study of association between coronary heart disease mortality rates in men and the relative frequencies of common allelic variations in the gene coding for apolipoprotein E.** *Hum Genet* 1998, **103**;234-241.

75. Merriman T, Twellis R, Merriman M, Eave I, Cox R, Cucca F, McKinney P, Shield J, Baum D, Bosi E *et al.*: **Evidence by allelic association-dependent methods for a type 1 diabetes polygene (IDDM6) on chromosome 18q21.** *Hum Mol Genet* 1997, **6**:1003-1010.

76. Valdes AM, McWeeney S, Thomson G: **HLA Class II DR-DQ amino acids and insulin-dependent diabetes mellitus: application of the haplotype method.** *Am J Hum Genet* 1997, **60**:717-728.

77. Cavalli-Sforza LL, Cavalli-Sforza F (Eds): *The Great Human Diasporas.* Reading: Addison-Wesley; 1995.

78. Szabo CI, King M-C: **Population genetics of BRCA1 and BRCA2.** *Am J Hum Genet* 1997, **60**:1013-1020.

79. Spielman RS, McGinnis RE, Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).** *Am J Hum Genet* 1993, **52**:506-516.

80. Terwilliger JD, Ott J: **A haplotype-based haplotype relative risk statistic.** *Hum Hered* 1992, **42**:337-346.

81. Kaplan NL, Martin ER, Weir BS: **Power studies for the transmission/disequilibrium test with multiple alleles.** *Am J Hum Genet* 1997, **60**:691-702.

82. Spielman RS, Ewens WJ: **A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test.** *Am J Hum Genet* 1998, **62**:450-458.

83. Curtis D: **Use of siblings as controls in case-control association studies.** *Ann Hum Genet* 1997, **61**:319-333.

84. Schaid DJ, Li H: **Genotype relative-risks and association tests for nuclear families with missing parental data.** *Genet Epidemiol* 1997, **14**:1113-1118.

85. Boehnke M, Langefeld CD: **Genetic association mapping based on discordant sib pairs: the discordant alleles test.** *Am J Hum Genet* 1998, **62**:950-961.

86. Excoffier L, Slatkin M: **Incorporating genotype of relatives into a test of linkage disequilibrium.** *Am J Hum Genet* 1998, **62**:171-180.

87. Sham P: **Transmission/disequilibrium tests for multiallelic loci.** *Am J Hum Genet* 1997, **61**:774-777.

88. Allison DB: **Transmission-disequilibrium tests for quantitative traits.** *Am J Hum Genet* 1997, **60**:676-690.

89. • Rabinowitz D: **A transmission/disequilibrium test for quantitative trait loci.** *Hum Hered* 1997, **47**:342-350.
The future of human gene mapping is undoubtedly going to focus more on quantitative variation and less on qualitative variation. To this end, the development of statistical methods, such as this one, to extend the current thinking to quantitative phenotypes is critical.

90. • Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE: **True and false positive peaks in genome-wide scans: applications of length-biased sampling to linkage mapping.** *Am J Hum Genet* 1997, **61**:430-438.
In this paper, relevant to association mapping, the general theory of the expected distribution of true and false positive shared segments is derived, and it is shown that the length of a haplotype is not so useful a measure of significance as the haplotype frequency in the population, as 25% of the false positives (assuming there are false positive IBD segments shared) will be longer than the true positives.

91. Martin ER, Kaplan NL, Weir BS: **Tests for linkage and association in nuclear families.** *Am J Hum Genet* 1997, **61**:439-448.

92. Teng J, Siegmund D: **Combining information within and between pedigrees for mapping complex traits.** *Am J Hum Genet* 1997, **60**:979-992.

93. • Trembath RC, Clough RL, Rosbotham JL, Jones AB, Camp RDR, Frodsham A, Browne J, Barber R, Terwilliger JD, Lathrop GM *et al.*: **Identification of a major susceptibility locus on chromosome 6p and evidence for further disease loci revealed by a two-stage genomewide search in psoriasis.** *Hum Mol Genet* 1997, **6**:813-820.
This paper applies a joint linkage and LD analysis to the MHC region in psoriasis pedigrees. The joint analysis of these phenomena in a rudimentary likelihood approach greatly increased the sensitivity of this analysis. Note that there are some unfortunate typographical problems where subscripts in formulas are written as L-D instead of as L_D, making this a tough read.

94. • Zhao LP, Aragaki C, Hsu L, Quiaoit F: **Mapping of complex traits by single-nucleotide polymorphisms.** *Am J Hum Genet* 1998, **63**:225-240.
Although it is unfortunate that the authors stress the issue of the diallelic single nucleotide polymorphisms in this manuscript, the general theory is a very useful means of combining linkage and association methods in a single statistical framework. This area of research will become more and more important with time, whether or not SNP markers become the markers of choice.

95. Camp NJ: **Genomewide transmission/disequilibrium testing – consideration of the genotype relative risks at disease loci.** *Am J Hum Genet* 1997, **61**:1424-1430.

96. •• Elston RC: **Algorithms and inferences: the challenge of multifactorial diseases.** *Am J Hum Genet* 1997, **60**:255-262.
In this paper, Elston's perspectives on the hot topics of the day are explained in a very clear and intuitive manner. While this paper focuses on linkage analysis in complex disease, it gets to the heart of some important philosophical issues about how to interpret the results of a linkage and by extension an association study from his own unique perspective.

97. Morton NE: **Significance levels in complex inheritance.** *Am J Hum*
•• *Genet* 1998, **62**:690-697.
In the field of human genetics, it would be irresponsible to highly recommend the treatises on interpretation of significance testing by Elston [96••]and Kruglyak [36••] and not to suggest just as strongly that one read Newton Morton's opinions on the subject. Especially in the area of association mapping, these insights are useful to understand in developing your own opinions about interpretation.

98. Vieland VJ, Hodge SE: **Review of 'Statistical evidence: a likelihood**
•• **paradigm' by Richard Royall.** *Am J Hum Genet* 1998, **63**:283-288.
This is an excellent review of an excellent book – it is strongly recommended that both the review and the book be read by anyone charged with interpreting results of an association study or linkage analysis. This review and book are further complements to the papers of Morton [97••], Elston [96••], and Kruglyak [36••] listed above, and is highly recommend reading for everyone. Read them all and make your own interpretations.

99. Staessen JA, Wang JG, Ginocchio G, Petrov V, Saavedra AP, Soubrier F, Vlietnick R, Fagard R: **The deletion/insertion polymorphism of the angiotensin converting enzyme gene and cardiovascular-renal risk.** *J Hypertens* 1997, **15**:1579-1592.

100. Goldgar DE, Easton DF: **Optimal strategies for mapping complex diseases in the presence of multiple loci.** *Am J Hum Genet* 1997, **60**:1222-1232.

101. Collins F, Galas D: **A new five-year plan for the US Human Genome Project.** *Science* 1993, **262**:43-46.

102. Collins F, Guyer M, Chakravarti A: **Variations on a theme: cataloging human DNA sequence variation.** *Science* 1997, **278**:1580-1581.

103. Ajioka RS, Jorde LB, Gruen JR, Yu P, Dimitrova D, Barrow J,
•• Radisky E, Edwards CQ, Griffen LM, Kushner JP: **Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes.** *Am J Genet* 1997, **60**:1439-1447.
This paper discusses the evolution of the hemochromatosis locus, and evaluates a number of different statistical methods for detection of LD and its application to estimation of the map position of a disease gene. In this case, a disease is analyzed for which almost all disease alleles in the population are identical by descent from a common ancestor. Multi-locus methods are emphasized.

104. Allamand V, Beckmann JS: **Mapping using linkage-disequilibrium**
• **estimates: a comparative study.** *Hum Hered* 1997, **47**:237-240.
Another comparison of statistical methods for the analysis of LD for a somewhat more complex genetic disorder than hemochromatosis. In this study, single locus disequilibrium measures are emphasized in a specific applied example from an interesting population dataset.

105. Wicking C, Shanley S, Smyth I, Gillies S, Negus K, Graham S, Suthers G, Haites N, Edwards M, Wainwright B *et al.*: **Most germ-line mutations in the nevoid basal cell carcinoma lead to a premature termination of the PATCHED protein, and no genotype-phenotype correlations are evident.** *Am J Hum Genet* 1997, **60**:21-26.

106. Tumer Z, Lund C, Tolshave J, Vural B, Tonnesen T, Horn N: **Identification of point mutations in 41 unrelated patients affected with Menkes disease.** *Am J Hum Genet* 1997, **60**:63-71.

107. Chamberlain ME, Ubagai T, Mudd SH, Levy HL, Chou JY: **Dominant inheritance of isolated hypermethioninemia is associated with a mutation in the human methionine adenosyltransferase 1A gene.** *Am J Hum Genet* 1997, **60**:540-546.

108. Gatalica B, Pulkkinen L, Li K, Kuokkanen K, Ryynänen M, McGrath J, Uitto J: **Cloning of the human type XVII collagen gene (COL17A1), and detection of novel mutations in generalized atrophic benign epidermolysis bullosa.** *Am J Hum Genet* 1997, **60**:352-365.

109. Friedman KJ, Leigh MW, Czarnecki P, Feldman GL: **Cystic fibrosis transmembrane-conductance regulator mutations among African Americans.** *Am J Hum Genet* 1998, **62**:195-196.

110. Tysoe C, Whittaker J, Xuereb J, Cairns NJ, Cruts M, van Broeckhoven C, Wilcock G, Rubinsztein DC: **A presenilin-1 truncating mutation is present in two cases with autopsy-confirmed early-onset Alzheimer disease.** *Am J Hum Genet* 1998, **62**:70-76.

111. Felbor U, Suvanto EA, Forsius HR, Eriksson AW, Weber BHF: **Autosomal recessive Sorsby fundus dystrophy revisited: evidence for dominant inheritance.** *Am J Hum Genet* 1997, **60**:57-62.

112. Yu C-E, Oshima J, Wijsman EM, Nakura J, Miki T, Piussan C, Matthews S, Fu Y-H, Mulligan J, Martin GM, Schellenberg GD, Werner's Syndrome Collaborative Group: **Mutations in the consensus helicase domains of the Werner's syndrome gene.** *Am J Hum Genet* 1997, **60**:330-341.

113. Lailioti MD, Mirotsou M, Buresi C, Peitsch MC, Rossier C, Ouazzani R, Baldy-Moulinier M, Bottani A, Malafosse A, Antonarakis SE: **Identification of mutations in Cystatin B, the gene responsible for the Univerricht-Lundborg type of progressive myoclonus epilepsy (EPM1).** *Am J Hum Genet* 1997, **60**:342-351.

114. Edwards SJ, Gladwin AJ, Dixon MJ: **The mutational spectrum in Treacher-Collins syndrome reveals a predominance of mutations that create a premature-termination codon.** *Am J Hum Genet* 1997, **60**:515-524.

115. De Paepe A, Nuytinck L, Hausser I, Anton-Lamprecht I, Naeyaert J-M: **Mutations in the COL5A1 gene are causal in the Ehlers-Danlos syndromes I and II.** *Am J Hum Genet* 1997, **60**:547-554.

116. Holm IA, Huang X, Kunkel LM: **Mutational analysis of the PEX gene in patients with X-linked hypophosphatemic rickets.** *Am J Hum Genet* 1997, **60**:790-797.

117. Akerman BR, Natowicz MR, Kaback MM, Loyer M, Campeau E, Gravel RA: **Novel mutations and DNA-based screening in non-Jewish carriers of Tay-Sachs disease.** *Am J Hum Genet* 1997, **60**:1099-1106.

118. Richard I, Brenguier RL, Dincer P, Roudaut C, Bady B, Burgunder J-M, Chemaly R, Garcia CA, Halaby G, Jackson CE *et al.*: **Multiple independent molecular etiology for limb-girdle muscular dystrophy type 2A patients from various geographical origins.** *Am J Hum Genet* 1997, **60**:1128-1138.

119. Puy H, Deybach JC, Lamoril J, Robreau AM, Da Silva V, Gouya L, Grandchamp B, Nordmann Y: **Molecular epidemiology and diagnosis of PBG deaminase see defects in acute intermittent porphyria.** *Am J Hum Genet* 1997, **60**:1373-1383.

120. Lin T, Orrison BM, Leahey A-M, Suchy SF, Bernard DJ, Lewis RA, Nussbaum RL: **Spectrum of mutations in the OCRL1 gene in the Lowe oculocerebrorenal syndrome.** *Am J Hum Genet* 1997, **60**:1384-1388.

121. Peral B, Gamble V, Strong C, Ong ACM, Sloane-Stanley J, Zerres K, Winearls CG, Harris PC: **Identification of mutations in the duplicated region of the polycystic kidney disease 1 gene (PKD1) by a novel approach.** *Am J Hum Genet* 1997, **60**:1399-1410.

122. Berg JN, Gallione CJ, Stenzel TT, Johnson DW, Allen WP, Schwartz CE, Jackson CE, Porteous MEM, Marchuk DA: **The activin receptor-like kinase 1 gene: genomic structure and mutations in hereditary hemorrhagic telangiectasia type 2.** *Am J Hum Genet* 1997, **61**:60-67.

123. Mancini D, Singh S, Ainsworth P, Rodenheiser D: **Constitutively methylated CpG dinucleotides as mutation hot spots in the retinoblastoma gene (RB1).** *Am J Hum Genet* 1997, **60**:80-87.

124. Munroe PB, Mitchison HM, O'Rawe AM, Anderson JW, Boustany R-M, Lerner TJ, Taschner PEM, de Vos N, Breuning MH *et al.*: **Spectrum of mutations in the Batten disease gene (CLN3).** *Am J Hum Genet* 1997, **61**:310-316.

125. Shah AB, Chernov I, Zhang HT, Ross BM, Das K, Lutsenko S, Parano E, Pavone L, Evgrafov O, Ivanova-Smolenskaya IA *et al.*: **Identification and analysis of mutations in the Wilson disease gene (ATP7B): population frequencies, genotype-phenotype correlation, and functional analyses.** *Am J Hum Genet* 1997, **61**:317-328.

126. Philippe C, Porter DE, Emerton ME, Wells DE, Hamish A, Simpson RW, Monaco AP: **Mutation screening of the EXT1 and EXT2 genes in patients with hereditary multiple exostoses.** *Am J Hum Genet* 1997, **61**:520-528.

127. Veldhuisen B, Saris JJ, de Haij S, Hayashi T, Reynolds DM, Mochizuki T, Elles R, Fossdal R, Bogdanova N, van Dijk MA *et al.*: **A spectrum of mutations in the second gene for autosomal dominant polycystic kidney disease (PKD2).** *Am J Hum Genet* 1997, **61**:547-555.

128. Buraczynska M, Wu W, Fujita R, Buraczynska K, Phelps E, Andreasson S, Bennett J, Birch DG, Fishman GA, Hoffman DR *et al.*: **Spectrum of mutations in the RPGR gene that are identified in 20% of families with X-linked retinitis pigmentosa.** *Am J Hum Genet* 1997, **61**:1287-1292.

129. Hovnanian A, Rochat A, Bodemer C, Petit E, Rivers CA, Prost C, Fraitag S, Christiano AM, Uitto J, Lathrop M *et al.*: **Characterization of 18 new mutations in COL7A1 in recessive dystrophic epidermolysis bullosa provides evidence for distinct molecular mechanisms underlying defective anchoring fibril formation.** *Am J Hum Genet* 1997, **61**:599-610.

130. Adato A, Weil D, Kalinski H, Pel-Or Y, Ayadi H, Petit C, Korostishevsky M, Bonne-Tamir B: **Mutation profile of all 49 exons of the human myosin VIIA gene, and haplotype analysis, in Usher 1B families from diverse origins.** *Am J Hum Genet* 1997, **61**:813-821.

131. Sampson JR, Maheshwar MM, Aspinwall R, Thompson P, Cheadle JP, Ravine D, Roy S, Haan E, Bernstein J, Harris PC: **Renal cystic disease in tuberous sclerosis: role of the polycystic kidney disease 1 gene.** *Am J Hum Genet* 1997, **61**:843-851.

132. Kikawa Y, Inuzuka M, Jin BY, Kaji S, Koga J, Yamamoto Y, Fujisawa K, Hata I, Nakai A, Shigematsu Y *et al.*: **Identification of genetic mutations in Japanese patients with fructose-1,6-biphosphatase deficiency.** *Am J Hum Genet* 1997, **61**:852-861.

133. Tsou HC, Teng DH-F, Ping XL, Brancolini V, Davis T, Hu R, Xie XX, Gruener AC, Schrager CA, Christiano AC *et al.*: **The role of MMAC1 mutations in early-onset breast cancer: causative in association with Cowden syndrome and excluded in BRCA1-negative cases.** *Am J Hum Genet* 1997, **61**:1036-1043.

134. Suzuki Y, Shirato S, Taniguchi F, Ohara K, Nishimaki K, Ohta S: **Mutations in the TIGR gene in familial primary open-angle glaucoma in Japan.** *Am J Hum Genet* 1997, **61**:1202-1203.

135. Savino M, Ianzano L, Strippoli P, Ramenghi U, Arslanian A, Bagnara GP, Joenje H, Zelante L, Savola A: **Mutations of the Fanconi anemia group A gene (FAA) in Italian patients.** *Am J Hum Genet* 1997, **61**:1246-1253.

136. Lynch ED, Ostermeyer EA, Lee MK, Arena JF, Ji HL, Dann J, Swisshelm K, Suchard D, MacLeod PM, Kvinnsland S *et al.*: **Inherited mutations in PTEN that are associated with breast cancer, cowden disease, and juvenile polyposis.** *Am J Hum Genet* 1997, **61**:1254-1260.

137. Aral B, Benelli C, Ait-Ghezala G, Amessou M, Fouque F, Maunoury C, Creau N, Kamoun P, Marsac C: **Mutations in PDX1, the human lipoyl-containing component X of the pyruvate dehydrogenase-complex gene on chromosome 11p1, in congenital lactic acidosis.** *Am J Hum Genet* 1997, **6**:1318-1326.

138. Schmidtchen A, Greenberg D, Zhao HG, Li HH, Huang Y, Tieu P, Zhao HZ, Cheng S, Zhao Z, Whitley CB *et al.*: **NAGLU mutations underlying Sanfilippo syndrome type B.** *Am J Hum Genet* 1998, **62**:64-69.

139. Mallery DL, Tanganelli B, Colella S, Steingrimsdottir H, van Gool AJ, Troelstra C, Stefanini M, Lehmann AR: **Molecular analysis of mutations in the CSB (ERCC6) gene in patients with Cockayne syndrome.** *Am J Hum Genet* 1998, **62**:77-85.

140. Bassett JHD, Forbes SA, Pannett AAJ, Lloyd SE, Christie PT, Wooding C, Harding B, Besser GM, Edwards CR, Monson JP *et al.*: **Characterization of mutations for patients with multiple endocrine neoplasia type I.** *Am J Hum Genet* 1998, **62**:232-244.

141. Mitchell GA, Ozand PT, Robert M-F, Ashmarina L, Roberts J, Gibson KM, Wanders RJ, Wang S, Chevalier I, Plochl E, Miziorko H: **HMG CoA lyase deficiency: identification of five causal point mutations in codons 41 and 42, including a frequent Saudi Arabian mutation, R41Q.** *Am J Hum Genet* 1998, **62**:295-300.

142. Briggs MD, Mortier GR, Cole WG, King LM, Golik SS, Bonaventure J, Nuytinck L, de Paepe A, Leroy JG, Biesecker L *et al.*: **Diverse mutations in the gene for cartilage oligomeric matrix protein in the pseudoachondroplasia-multiple epiphyseal dysplasia disease spectrum.** *Am J Hum Genet* 1998, **62**:311-319.

143. Bejjani BA, Lewis RA, Tomey KF, Anderson KL, Dueker DK, Jabak M, Astle WF, Otterud B, Leppert M, Lupski JR: **Mutations in CYP1B1, the gene for cytochrome P4501B1, are the predominant cause of primary congenital glaucoma in Saudi Arabia.** *Am J Hum Genet* 1998, **62**:325-333.

144. Daniele A, Parenti G, d'Addio M, Andria G, Ballabio A, Meroni G: **Biochemical characterization of arylsulfatase E and functional analysis of mutations found in patients with X-linked chondrodysplasia punctata.** *Am J Hum Genet* 1998, **62**:562-572.

145. Oh J, Ho L, Ala-Mello S, Amato D, Armstrong L, Bellucci S, Carakushansky G, Ellis JP, Tong C-T, Green JS *et al.*: **Mutation analysis of patients with Hermansky-Pudlak syndrome: a frameshift hot spot in the HPS gene and apparent locus heterogeneity.** *Am J Hum Genet* 1998, **62**:593-598.

146. Beltran-Valero de Bernabe D, Granadino B, Chiarelli I, Porfirio B, Mayatepek E, Aquaron R, Moore MM, Festen JJM, Sanmarti R, Penalva MA *et al.*: **Mutation and polymorphism analysis of the human homogentisate 1,2-dioxygenase gene in alkaptonuria patients.** *Am J Hum Genet* 1998, **62**:776-784.

147. Burwinkel B, Bakker HD, Herschkvitz E, Moses SW, Shin YS, Kilimann MW: **Mutations in the liver glycogen phosphorylase gene (PYGL) underlying glycogenesis type VI (Hers disease).** *Am J Hum Genet* 1998, **62**:785-791.

148. Kelley PM, Harris DJ, Comer BC, Askew JW, Fowler T, Smith SD, Kimberling WJ: **Novel mutations in the connexin 26 gene (GJB2) that cause autosomal recesive (DFNB1) hearing loss.** *Am J Hum Genet* 1998, **62**:792-799.

149. Schnur RE, Gao M, Wick PA, Keller M, Benke PJ, Edwards MJ, Grix AW, Hockey A, Jung JH, Kidd KK *et al.*: **OA1 mutations and deletions in X-linked ocular albinism.** *Am J Hum Genet* 1997, **62**:800-809.

150. Conley ME, Mathias D, Treadaway J, Minegishi Y, Rohrer J: **Mutations in Blk inpatients with presumed X-linked agammaglobulinemia.** *Am J Hum Genet* 1998, **62**:1034-1043.

151. Thony B, Neuheiser F, Kierat L, Blaskovics M, Arn PH, Ferreira P, Rebrin I, Ayling J, Blau N: **Hyperphenylalaninemia with high levels of 7-biopterin is associated with mutations in the PCBD gene encoding the bifunctional protein pterin-4a-carbinolamine dehydratase and transcriptional coactivator (DCoH).** *Am J Hum Genet* 1998, **62**:1302-1311.

152. Vargas-Poussou R, Feldmann D, Vollmet M, Konrad M, Kely L, van den Heuvel LPWJ, Tebourbi L, Brandis M, Karolyi L, Hebert SC *et al.*: **Novel molecular variants of the Na-K-2Cl cotransporter gene are responsible for antenatal bartter syndrome.** *Am J Hum Genet* 1998, **62**:1332-1340.

153. Rufenacht UB, Gouya L, Schneider-Yin X, Puy H, Schaffer BW, Aquaron R, Nordmann Y, Minder EI, Deybach JC: **Systematic analysis of molecular defects in the ferrochelatase gene from patients with erythropoetic protoporphyria.** *Am J Hum Genet* 1998, **62**:1341-1352.

154. Malzac P, Webber H, Moncla A, Graham JM, Kuklich M, Williams C, Pagon RA, Ramsdell LA, Kishino T, Wagstaff J: **Mutation analysis of UBE3A in Angelma syndrome patients.** *Am J Hum Genet* 1998, **62**:1353-1360.

155. Krantz ID, Colliton RP, Genin A, Rand EB, Li L, Piccoli DA, Spinner NB: **Spectrum and frequency of Jagged1 (JAG1) mutations in Alagille syndrome patients.** *Am J Hum Genet* 1998, **62**:1361-1369.

156. Paznekas WA, Cunningham ML, Howard TD, Korf BR, Lipson MH, Grix AW, Feingold M, Goldberg R, Borochowitz Z, Aleck K *et al.*: **Genetic heterogeneity of Saethre-Chotzen syndrome, due to TWIST and FGFR mutations.** *Am J Hum Genet* 1998, **62**:1370-1380.