

Finding genes underlying risk of complex disease by linkage disequilibrium mapping

Andrew G Clark

Identification of genes that harbor variation associated with inter-individual differences in risk of complex diseases remains one of the most challenging and important problems in human genetics. For genetic variants that are sufficiently common and have sufficiently large effects, direct tests of association through linkage disequilibrium with anonymous SNPs may prove effective. But the two critical parameters — the frequency of risk-inflating alleles and the magnitudes of their effect on risk — remain largely unknown. In this review we consider the latest information regarding the likely efficacy of the linkage disequilibrium mapping approach.

Addresses

Molecular Biology and Genetics, 107 Biotechnology Building,
Cornell University, Ithaca, New York 14853, USA
e-mail: ac347@cornell.edu

Current Opinion in Genetics & Development 2003, **13**:296–302

This review comes from a themed issue on
Genetics of disease
Edited by James R Lupski and Lap-Chee Tsui

0959-437X/03/\$ – see front matter
© 2003 Elsevier Science Ltd. All rights reserved.

DOI 10.1016/S0959-437X(03)00056-X

Abbreviations

ASP	affected sib pair
HapMap	Haplotype Map
LD	linkage disequilibrium
NIH	National Institutes of Health
SNP	single nucleotide polymorphism

Introduction

Health problems that appear to aggregate within families but that do not segregate like a simple Mendelian gene pose a special problem for researchers trying to either predict the risk of the disorder or to identify relevant genes for understanding etiology. Traditionally, linkage methods have served extremely well for Mendelian disorders, and the same approach of fitting linkage models to pedigree-structured data in which phenotypes and marker genotypes are scored has been reasonably effective for complex traits as well. The problem is that pedigree methods suffer from the fact that the resolution of the mapping depends on both sample size and marker density, and even the largest studies typically have a rather poor resolution of 5–10 cM. More recently it was discovered that one can apply similar analysis to affected sib pairs, noting that sharing of marker alleles and of phenotypes is more likely when the marker is

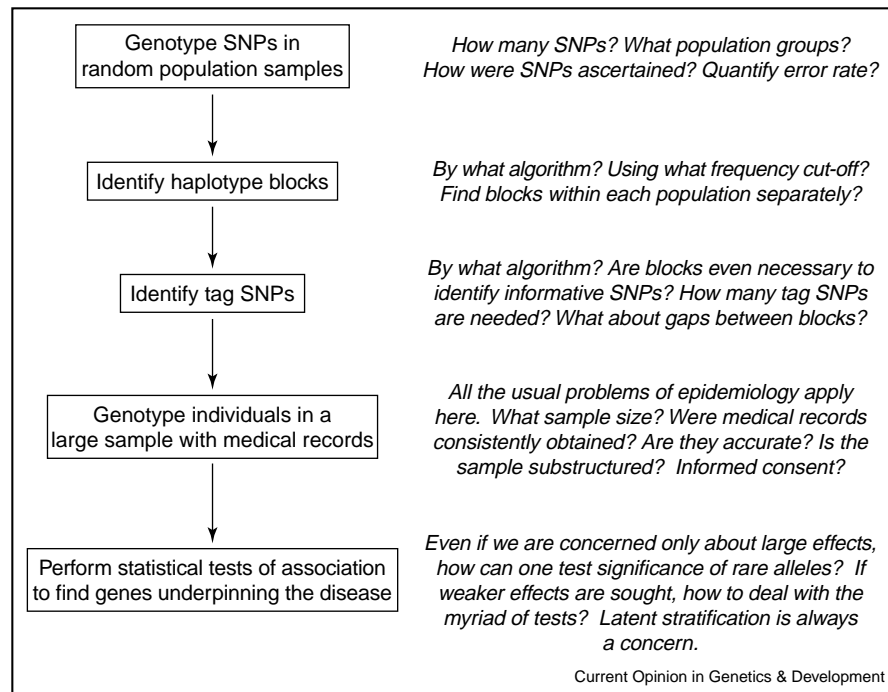
closely linked to segregating variation that causes trait variation. Although affected sib pair (ASP) methods have the big advantage that much larger samples can be obtained, they lack the advantage of acquiring information about linkage phase that a multigeneration pedigree provides, and so in the end the resolution of ASP methods also are less than optimal. The final assessment of the efficacy of using whole genome linkage disequilibrium (LD) scans to find genes associated with risk of complex disease will have to wait until the approach is actually tried. In the meantime, I here discuss recent work that has been done seeking to improve the chances for success of the method by characterizing and analyzing the haplotype structure of human genetic variation.

Directly testing disease association

Risch and Merikangas [1] made the observation that an outbreeding population has some properties like a large extended family — namely there are many meioses in which the association between a marker and a disease-associated allele can recombine. But if the marker and disease-associated alleles are found to be in tight statistical association, this may amount to evidence that they are closely linked. There is a large body of theory behind this notion, and the theory describes many factors that influence the magnitude and structure of LD in the genomes of a population.

The paper of Risch and Merikangas [1] marked the start of an explosion of interest in using direct statistical association between markers and diseases for purposes of gene mapping (Figure 1). They considered the best-case scenario of a disease caused by genes that may influence disease risk across all other genetic backgrounds, and that have relatively high penetrance. LD mapping was first applied to simple Mendelian disorders in populations that had resulted from expansion from relatively few founders [2]. The disease-causing allele was presumed to have occurred uniquely in the population on a single haplotype background, and over time recombination would erode the size of this initial haplotype that remains intact and associated with the disease. Hill and Weir [3] then asked how well one might map a gene in an equilibrium population, and the results were not nearly as promising. But these early efforts made clear that many parameters affect the success of LD mapping — including marker density, sample size, disease incidence, allele penetrance, allele frequency, population subdivision, and past demographic history of the population. Note that these are all issues that need to be considered, even if the disease were caused by a single

Figure 1



The steps being taken to make use of genome-wide patterns of LD for identification of genes underlying complex diseases. For each step, there are a number of unanswered questions, listed to the right. In many cases, these seem like surmountable technical issues, in others, there may be many solutions, including some quite different from those proposed to date that will prove to be effective. And for some of these questions, we simply do not know how large the hurdle will be.

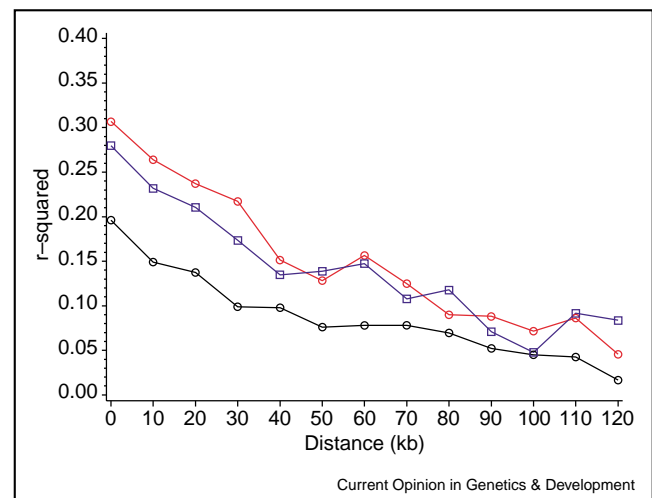
gene, and when complications of environmental effects and genotype–environment interactions are added, one begins to see why these are called complex diseases.

Linkage disequilibrium across the genome

To assess the overall efficacy and cost of LD mapping, we first need to determine the distribution of spans of the human genome that exhibit LD. This had been done for several human genes by resequencing to obtain many single nucleotide polymorphisms (SNPs) in the same gene [4–7]. Figure 2 shows the relationship between physical separation between SNPs and two common metrics for LD. Other metrics for LD are evaluated by Devlin and Risch [8]. One problem with this approach is that from one study to another, different sample sizes from different populations were genotyped at different numbers and densities of SNPs, making it hard to contrast regions of the genome.

A more complete picture of the landscape of LD across the human genome has also been obtained by genotyping SNPs at multiple local genomic regions in the same sample ([9,10]; AG Clark *et al.*, unpublished data). This approach shows unambiguously that the rate of decay of LD varies widely among regions of the genome, and that there is a significant negative correlation between local

Figure 2



Plot of the decay of average LD (measured as r^2) versus the physical separation of SNPs. This plot was done using genotypes from the SNP Consortium panel used to produce a linkage map and a map of LD (AG Clark *et al.*, unpublished data). The red line is for the Asian sample, the blue for European American, and the black for African American. Note the lower LD in the African-American sample, and the average tendency for LD to have halved in ~60 kb. This figure varies widely across different regions of the genome.

rate of recombination (inferred from pedigrees) and local LD. In the regions of low recombination, LD extending >500 kb was not uncommon. Another approach is to consider SNPs across entire chromosomes. An analysis of 1504 SNPs spaced at an average interval of 15 kb across chromosome 22 revealed remarkable heterogeneity in LD, with stretches up to 804 kb in nearly complete LD in the CEPH families [11]. A more extensive analysis of 19,860 SNPs on chromosomes 6, 21, and 22 showed that less than half of the genome fell into any of the published block definitions, but that for those regions that had sufficient LD to be considered block-like, the power to detect association might be reasonably good (F De La Vega *et al.*, personal communication). Genome-scanning by LD mapping appears to be feasible today, with the caveat that one is only covering a subset of the genome that is in relatively high LD. To span the complete genome will require vastly more SNPs, and the exact figure will vary widely across populations.

Population subdivision, demography and linkage disequilibrium

There is a long history of interest in inferring the degree of population subdivision from genetic data, and application of this analysis to human genetic markers reveals that ~8–10% of the genetic variance is found within population groups [12,13]. When making inferences of association between genes and complex diseases, the need to understand population subdivision is critically important. If one does a case-control study, and the samples under study are a mix of two somewhat isolated population groups — one with high disease incidence and one with low incidence — then there will be a spurious association between this disease and any genetic marker that shows allele frequency differences between the population strata. The need to understand population structure and methods for dealing with it in case/control designs is reviewed by Jorde [14] and by Pritchard and Donnelly [15].

In addition to introducing differences in allele frequency, partial subdivision of populations (i.e. departure from universal random mating) can also result in haplotype frequencies varying among geographic regions. If haplotypes differ in frequency, then the patterns of LD may also be heterogeneous. Empirical results are now showing that not only is this so, but that there are marked trends in the levels and extent of LD across human populations. In particular, there seems to be less LD within African populations than in populations outside of Africa [9*,16]. The primary reason is that human migrations out of Africa probably only sampled a subset of the total diversity that was within Africa, and the resulting founder effect could have inflated LD [17]. Past human demography included population founding events, expansion, and migration, and each of these factors plays a complex role in determining local patterns of LD [18*].

Complex disorders are not simple

Even if LD mapping of single genes were simple, mapping complex traits is an enormous challenge for the same reasons that it is so difficult to draw firm conclusions from epidemiological data. Genetic variation is likely to contribute to overall risk of many complex diseases, but the genetic component may be small compared to some environmental insults, and the fact that genes and environment interact, and that health is something that is deeply context dependent (CF Sing, JH Stengård, SLR Kardia, personal communication), means that we need to be acutely aware of these complexities before being too confident that simple methods will yield any meaningful results. All of the challenges of classical quantitative genetics [19] are with us in complex disease research, along with the major limitation of being unable to replicate genotypes in experimental manipulations. That a given disorder, such as schizophrenia, might be caused by multiple totally different sets of genetic and environmental conditions does not make it any easier.

The field of epidemiology seeks to understand complex disease causation by appropriate stratification of the sample. The need to retain an appreciation for the epidemiological perspective has been argued for analysis of complex diseases [20], in light of the complex causation that is implied [21]. A frustration in this field is that much of the analysis that is done to determine statistical power and to determine the number of SNPs needed for a given resolution of mapping assumes that the disease is caused by a single Mendelian gene with appreciable penetrance. The challenge is not the determination of how many SNPs will be needed to map a Mendelian trait — that is a simple problem compared to the issue of dealing with diseases in which perhaps every case has a unique chain of genetic and environmental causal elements [21–23].

Models for the genetics of complex disorders

There is a long history in genetic analysis that points to the power of a good model. If we formulate a scheme whereby genes affect a trait, we are much more able to test and either reject or accept the model, compared to a more open-ended situation. Key parameters in whole-genome association testing are the number of genes that are having a causal effect on risk, the frequency of the variant alleles, and the magnitudes of effect of those alleles on risk. Before we consider the complexities of dealing with the genetics of the traits themselves, there is much modeling to be done simply on the genetic structure of the markers. In a population undergoing mutation and random genetic drift, population genetics theory can provide a complete description of the expected frequencies of polymorphic nucleotides, and of the relationships among alleles sampled from the population [24]. The neutral coalescent has been a remarkably powerful mathematical construct for modeling the population genetics of markers, and recent extensions

allow it to cover the case of intragenic recombination [25] and population growth. These models have been useful in early efforts to infer what likely genome-wide patterns of LD will be, and to estimate the number of SNPs needed as markers to cover the full genome in LD maps.

Unfortunately, our confidence in understanding how markers behave is not matched by our confidence in understanding the genetics underlying complex traits. If the SNPs that are actually causal to disease risk are no different from neutral SNPs, then we know that the predicted frequency spectrum will have a large number of very rare SNPs [26]. But recently there has been hope expressed that many common diseases will have relatively common alleles as a cause, and reasons to support this 'common disease/common variant' hypothesis have been offered [27]. Although it would certainly assist in our efforts to find genes underlying complex disease if the common disease/common variant hypothesis were true, and undoubtedly some disorders will have risk-increasing alleles at appreciable frequency, it is important to carefully consider how reasonable it is to expect that common disease/common variant will be the rule. Pritchard [28*] and Pritchard and Cox [29] marshal the evidence and the models, and show that even in a population with founder effects and growth, and even with late-onset diseases where selection against causal alleles may be weak [30,31], on balance the effect of mutation-selection balance is to drive down the frequency of alleles that reduce fitness. If those alleles that cause disease were in the past beneficial in some way, such as to reduce risk from malaria (like β^S globin or G6PD deficiency), then of course the alleles could have been driven to high frequency by past selection. The fact that we can list alleles of this type is a bit misleading, because these are the associations that are easiest to detect by virtue of the commonness of the alleles.

If the alleles associated with complex disease are generally very rare (frequency <1%), then the standard ways that we are considering to map them with LD will not be very effective. In fact, mapping the very rare alleles will be a challenge by any method. One approach that works well, even with rare alleles, is to perform linkage analysis on pedigrees. If a disease is caused by one gene with many distinct and individually rare alleles, then a large pedigree study can still detect the association by virtue of co-transmission of flanking markers. This means that any method that makes use of transmission in families is likely to perform much better in the context of rare alleles. Methods such as the transmission disequilibrium test [32] fit this description. Other methods, such as those that make explicit use of demographic patterns like admixture mapping [33], also show promise. But all these methods bring with them trade-offs. Pedigree approaches may allow ascertainment of rare alleles, and identification of linkage to genes with rare deleterious alleles segregating, but their effectiveness is seriously compromised if the disorder is

genetically heterogeneous. Because sample sizes of pedigree studies are much smaller than a random population sample could be, pedigree approaches are limited to detection of alleles of relatively large effects.

Selection in the human genome

A factor that inflates LD in the human genome more directly and strongly than any other is natural selection. This is especially evident in cases where a single gene has an influence on the risk from a disease, such as the improved resistance to Vivax malaria by people with the Duffy null allele [34] or increased resistance to *Plasmodium malariae* in individuals with the low-activity alleles of *G6PD* [35]. The recent generation of near genome-wide datasets on SNP genotypes has opened the possibility of doing genome-wide screens for the impact of past selection. The idea is to perform any of a welter of tests for selection that have appeared in the literature [36–40] either to each gene or to sliding windows for the whole genome. Initial reports suggest that the evidence for natural selection in the human genome may be widespread [41,42], and this in turn means that the true pattern of LD may be quite different from that predicted by present modeling efforts. Already the chromosome-wide surveys of LD show that levels of LD vary much more widely than expected by chance.

Why HapMap?

The NIH Haplotype Map (HapMap) project is the largest single project in human population genetics ever attempted, and as a result it has received some harsh criticism. As of writing, the exact scope of the project is unclear, but it will entail a large quantity of SNP genotyping in several human population groups. Given that the project will be completed and the genotype data will be collected, the constructive challenge we face is to formulate the best questions and the best use of the resulting data. One need not fully accept the idea that the genome is broken cleanly into discrete haplotype blocks [10,43*,44*,45], nor that these blocks necessarily imply that there are recombination hotspots that separate them [46*]. Haplotype blocks may cover as little as one-third of a chromosome, so even if an optimal algorithm for identifying blocks can be agreed upon, there will still be the problem of finding methods to test SNPs in non-block regions. The plan is to collect the primary data with adequate quality control, and these data should be useful for revealing a great deal about past demographic history of human populations [17]. Our genome-wide studies of LD have been at too small a scale, both in numbers of SNPs and in sample size, to allow much more than a rudimentary effort to identify regions of the genome that have unusually high or low LD, or that may have faced a past selective sweep [41]. In the regions of the genome where dense SNP genotype data do exist, there are already sharp departures between theory and observation. In particular, it appears that the level of LD is lower than

expected at short physical scales. A likely explanation for the departure is that gene conversion serves to erode local LD but not so much long range LD [47*,48].

Another use of the HapMap data are to devise and apply strategies to use a subset of SNPs for association testing. The excitement over haplotype blocks is justified by the great need to avoid having to genotype millions of SNPs for a whole genome LD scan. If there were a way to use the local LD structure to select an optimal subset of SNPs that captured essentially the same information, then the HapMap project could result in savings of time and money (at least if one ignores the price of the HapMap project itself!) The observation that runs of SNPs seem to occur in LD, suggests that these segments of the chromosome may have remained intact without recombination for a longer period than flanks. By selecting a set of 'tagging' SNPs within each such block [49], one may have a means to reduce the genotyping cost. In summary, the HapMap project has the potential to be a valuable resource for the human genetics community, provided the sampling covers an adequate distribution of population groups, with adequate sample size and adequate numbers of SNPs, that the mode of ascertainment of the SNPs is carefully recorded, and that the complete genotype information on each individual is made freely available to the research community.

Several methods for inferring local blocks of sites that show little evidence for recombination (haplotype blocks) have been proposed. Some rely only on pairwise LD and are thus not strictly speaking identifying robust multisite haplotypes. If one knew the phasing of the multiple heterozygous SNPs in each individual, one would know the pair of haplotypes in that individual. Direct determination of haplotype phase — by allele-specific PCR, cloning, or other molecular means — adds a considerable burden to SNP genotyping, so it is rarely done. Fortunately, inference of haplotype phase from population samples can be done with reasonable accuracy, provided the sample size is sufficiently large and the rate of intra-region recombination is sufficiently low [50]. Statistical inference of haplotype phase is an area of active research, and a variety of Bayesian [51*,52,53] and graph theoretic [54] approaches are now available. Despite this interest, it remains untested exactly how much power is gained by application of association tests to data with inferred haplotype phase as opposed to testing association directly with unphased data. If knowing the haplotype phasing makes little difference, then it is possible that considerable gains in efficiency could be obtained by pooling DNA from multiple cases [55], although such pooling strategies come at a cost in reduced information about each sample.

Conclusions

The potential for a disease to be determined by a vast array of extremely rare alleles in many different genes

embedded in a network of highly epistatic genes with strong context-dependent environmental effects makes it possible to imagine that some diseases may have a genetic component but be truly unyielding by the proposed methods. But even in this worst-case scenario, we already know that not all complex diseases are this ill behaved, so the problem can be restated as finding efficient means to identify which diseases will allow progress by LD mapping [22]. A recent review of efforts to test associations stresses the role of meta-analysis to validate findings [56,57]. The National Institutes of Health HapMap project is somewhat controversial largely because of the uncertainty in the effectiveness of whole-genome LD mapping. Despite this uncertainty, large-scale efforts to directly test associations between genes and diseases in unrelated samples are underway in both the public and private sector. It would seem, then, that we will soon have a direct assessment of the efficacy of LD mapping for finding genes that underlie complex disorders.

Acknowledgement

This work was supported by grant HG02352 from the United States National Institutes of Health.

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. Risch N, Merikangas K: **The future of genetic studies of complex human diseases**. *Science* 1996, **273**:1516-1517.
 2. Hastbäckä J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E: **Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland**. *Nat Genet* 1992, **2**:204-211.
 3. Hill WG, Weir BS: **Maximum-likelihood estimation of gene location by linkage disequilibrium**. *Am J Hum Genet* 1994, **54**:705-714.
 4. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård 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, **63**:595-612.
 5. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome**. *Nat Rev Genet* 2002, **3**:299-309.
 6. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: **Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism**. *Am J Hum Genet* 2000, **67**:881-900.
 7. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengård J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF: **DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene**. *Nat Genet* 1998, **19**:233-240.
 8. Devlin B, Risch N: **A comparison of linkage disequilibrium measures for fine-scale mapping**. *Genomics* 1995, **29**:311-322.
 9. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES: **Linkage disequilibrium in the human genome**. *Nature* 2001, **411**:199-204. This paper took a broader look at LD in the human genome by selecting 19 sets of SNPs spanning windows of ~160 kb, and showed clearly that different regions had dramatically different rates of LD decay, and that different populations also had different rates of LD decay.

10. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al.*: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225-2229.
11. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibbling T, Tinsley E, Kirby S *et al.*: **A first-generation linkage disequilibrium map of human chromosome 22.** *Nature* 2002, **418**:544-548.
12. Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: **An apportionment of human DNA diversity.** *Proc Natl Acad Sci USA* 1997, **94**:4516-4519.
13. Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G: **Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms.** *Genome Res* 2002, **12**:602-612.
14. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10**:1435-1444.
15. Pritchard JK, Donnelly P: **Case-control studies of association in structured or admixed populations.** *Theor Popul Biol* 2001, **60**:227-237.
16. Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO *et al.*: **Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations.** *Am J Hum Genet* 2000, **66**:1882-1899.
17. Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M *et al.*: **Sequence variations in the public human genome data reflect a bottlenecked population history.** *Proc Natl Acad Sci USA* 2003, **100**:376-381.
18. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69**:1-14.
An assessment of the implications of observations of human LD and how well the observations correspond to theoretical population genetic predictions.
19. Barton NH, Keightley P: **Understanding quantitative genetic variation.** *Nat Rev Genet* 2002, **3**:11-21.
20. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405**:847-856.
21. Strohman R: **Maneuvering in the complex path from genotype to phenotype.** *Science* 2002, **296**:701-703.
22. Weiss KM, Terwilliger JD: **How many diseases does it take to map a gene with SNPs?** *Nat Genet* 2000, **26**:151-157.
23. Weiss KM, Clark AG: **Linkage disequilibrium and the mapping of complex human traits.** *Trends Genet* 2002, **18**:19-24.
24. Tajima F: **Evolutionary relationship of DNA sequences in finite populations.** *Genetics* 1983, **105**:437-460.
25. Nordborg M, Tavaré S: **Linkage disequilibrium: what history has to tell us.** *Trends Genet* 2002, **18**:83-90.
26. Ewens WJ: **The sampling theory of selectively neutral alleles.** *Theor Popul Biol* 1972, **3**:87-112.
27. Reich DE, Lander ES: **On the allelic spectrum of human disease.** *Trends Genet* 2001, **17**:502-510.
28. Pritchard JK: **Are rare variants responsible for susceptibility to complex disease?** *Am J Hum Genet* 2001, **69**:124-137.
A well-articulated argument for why it is likely that rare alleles will account for a substantial portion of the risk of complex diseases in humans.
29. Pritchard JK, Cox NJ: **The allelic architecture of human disease genes: common disease — common variant... or not?** *Hum Mol Genet* 2002, **11**:2417-2423.
30. Partridge L, Barton NH: **Optimality, mutation and the evolution of ageing.** *Nature* 1993, **362**:305-311.
31. Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H: **A polygenic basis for late-onset diseases.** *Trends Genet* 2003, **19**:97-106.
32. 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.
33. McKeigue PM, Carpenter JR, Parra EJ, Shriver MD: **Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations.** *Ann Hum Genet* 2000, **64**:171-186.
34. Hamblin M, Thompson EE, Di Rienzo A: **Complex signatures of natural selection at the Duffy blood group locus.** *Am J Hum Genet* 2002, **70**:369-383.
35. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J *et al.*: **Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance.** *Science* 2001, **293**:455-462.
36. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123**:585-595.
37. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133**:693-709.
38. Hudson RR, Kreitman M, Aguadé M: **A test of neutral molecular evolution based on nucleotide data.** *Genetics* 1987, **116**:153-159.
39. McDonald JH, Kreitman M: **Adaptive protein evolution at the *Adh* locus in *Drosophila*.** *Nature* 1991, **351**:652-654.
40. Yang Z: **PAML: a program for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**:555-556.
41. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res* 2002, **12**:1805-1814.
42. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome.** *Genetics* 2001, **158**:1227-1234.
43. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: **High-resolution haplotype structure in the human genome.** *Nat Genet* 2001, **29**:229-232.
One of the first papers to establish the existence and importance of haplotype blocks.
44. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP *et al.*: **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.** *Science* 2001, **294**:1719-1723.
A stunning demonstration of sequencing by hybridization, done on human-rodent cell hybrids that possess only a single human chromosome 21, so that all the data from a single cell line are truly representative of one haplotype.
45. Cardon LR, Abecasis GR: **Using haplotype blocks to map human complex trait loci.** *Trends Genet* 2003, **19**:135-140.
46. Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, Ankeney WM, Alfisi SV, Kuo FS *et al.*: **Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots.** *Nat Genet* 2003, **33**:382-387.
This paper provides a thoughtful analysis of the LD and haplotype structure inferred from SNPs on chromosome 19. The authors find that only about one-third of the chromosome is covered by haplotype blocks, and simulations show that there is no reason to have to invoke recombination hotspots in order to explain the observed blocks.
47. Przeworski M, Wall JD: **Why is there so little intragenic linkage disequilibrium in humans?** *Genet Res* 2001, **77**:143-151.
This paper attempts to fit the pattern of decay of LD with physical distance in the human genome to population genetic models and finds an important discrepancy. The observed decay in LD is faster than expected (based on direct estimates of recombination rate), especially at short distances. Possible explanations — including gene conversion, population growth, and hypermutation of CpG dinucleotides — all fail to fully accommodate the observed decay in LD.
48. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A: **Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels.** *Am J Hum Genet* 2001, **69**:831-843.

49. Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F *et al.*: **Haplotype tagging for the identification of common disease genes.** *Nat Genet* 2001, **29**:233-237.
50. Clark AG: **Inference of haplotypes from PCR-amplified samples of diploid populations.** *Mol Biol Evol* 1990, **7**:111-122.
51. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68**:978-989.
This is the paper that describes the algorithm behind PHASE, one of the popular haplotype inference programs in wide use. One of the more elegant applications of Bayesian inference in modern statistical genetics.
52. Niu T, Qin ZS, Xu X, Liu JS: **Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **70**:157-169.
53. Qin ZS, Niu T, Liu JS: **Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms.** *Am J Hum Genet* 2002, **71**:1242-1247.
54. Gusfield D: **Inference of haplotypes from samples of diploid populations: complexity and algorithms.** *J Comput Biol* 2001, **8**:305-323.
55. Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, Hollstein P, Boehnke M, Collins FS: **High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools.** *Proc Natl Acad Sci USA* 2002, **99**:16928-16933.
56. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: **A comprehensive review of genetic association studies.** *Genet Med* 2002, **4**:45-61.
57. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: **Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.** *Nat Genet* 2003, **33**:177-182.