

Population genomics: Linkage disequilibrium holds the key

David B. Goldstein* and Michael E. Weale†

Current efforts to find disease-causing genes depend on patterns of linkage disequilibrium in human populations. Recent work has shown that linkage disequilibrium can extend over much larger genomic regions than expected, and that the patterns of linkage disequilibrium can differ markedly among populations.

Addresses: *Galton Lab, Department of Biology University College London, 4 Stephenson Way, London NW1 2HE, UK. †Genostics Ltd, 28/30 Little Russell Street, London WC1A 2HN, UK.
E-mail: d.goldstein@ucl.ac.uk

Current Biology 2001, 11:R576–R579

0960-9822/01/\$ – see front matter
© 2001 Elsevier Science Ltd. All rights reserved.

In many ways, the Human Genome Project has been more successful than even its supporters might have expected. It has come in ahead of schedule with its flagship goal of providing a draft human sequence, and the genetic and physical maps generated have been put to spectacular use in identifying the genetic bases of scores of single-locus diseases. For these genetically simple diseases it is possible to determine the rough genomic position of the causal mutation by comparing the co-inheritance of variable marker loci and the disease through affected pedigrees with a methodology, linkage analysis, which is now routine. Unfortunately, the common diseases responsible for the vast majority of mortality and morbidity in the developed world are anything but simple.

Most cancers and cardiovascular, neuro-psychiatric, respiratory and infectious diseases are influenced by variation at multiple loci and show complicated dependence on environmental factors. This complexity is also reflected in large inter-individual variation in response to therapeutic treatment. As adverse drug reactions are responsible for more than 100,000 deaths each year in the US alone [1], variable drug reaction itself ranks as one of the primary challenges of contemporary biomedical research. In short, it is no time for triumphalism. The greatest challenges in human genetics remain ahead.

Population-based approaches

While linkage analysis may still have a part to play to meet these challenges, a promising alternative approach turns from families to populations, and from linkage analysis to association studies. As illustrated in the recent paper by Eric Lander, David Reich and colleagues [2], the shift to association studies makes the description of genetic variation in human populations a pre-requisite for the development of effective mapping strategies. The basic approach

used in an association study is straightforward. For example, to test the involvement of a single nucleotide polymorphism (SNP) in a specific condition, allele frequencies are compared in affected and un-affected individuals (or cases and controls). One advantage of a population-based association study, particularly in a case-control design, is that individuals can be selected to match environmental factors, such as age or lifestyle, that may also be important in the disease. Another is that risks can be properly assessed against the genetic and environmental 'background' of the population under study.

If we could test all candidate sites in this way, it would be possible to identify those variants that influence both the common diseases and variable drug reactions. In fact, simple theoretical calculations indicate that an allele resulting in a relative risk of two could be identified in this framework [3]. But we are a long way from knowing all the SNPs, even in any given population. Important subsets of SNPs will appear sooner, for example all coding SNPs. Even for such a restricted set, however, exhaustive typing in large case-control studies would currently be prohibitively expensive as a routine procedure.

Linkage disequilibrium in association studies

So if exhaustive typing is currently prohibitive, technically and economically, in most situations, what can be done? We have noted that SNP frequencies will differ if the SNP influences the trait. But a SNP variant could also be associated with the condition not because it is biologically causal, but because it is statistically correlated with a causal variant. This possibility arises because alleles at different loci are sometimes found together more or less often than expected based on their frequencies. In population genetics, this non-random pattern is called linkage disequilibrium.

To see how linkage disequilibrium could be used to map disease genes, consider a marker locus M , with two alleles M_1 and M_2 , and an unknown causative locus B , with one allele B_1 that is a risk factor for high blood pressure relative to the other allele B_2 . We expect B_1 to be elevated in cases relative to controls. Now imagine that the M and B loci are in disequilibrium, and specifically that the M_1 allele is more often found with the B_1 allele than the B_2 allele. In that case, not only B_1 is elevated in cases, but M_1 is too because of its association with B_1 . Thus if we did not know about the B locus, but typed the M locus, it could lead us to the B locus. But what exactly does such an association indicate? And how should we distribute markers through the genome in order to find causal variants?

Linkage disequilibrium in human populations

Lander and colleagues [2] have taken an important step toward addressing these questions. While there have been a number of studies of linkage disequilibrium in the past several years, most of them focused on relatively few markers in only one or a few genomic regions, limiting the utility of the results. Lander and colleagues [2] have taken the first post-genome approach, assessing associations at uniformly spaced intervals across 160 kilobases (kb) in each of 19 genomic regions, in samples from three populations. Presumably to match conditions to mapping studies, Lander and colleagues [2] used a 'core' coding SNP to anchor each of the 19 regions. To identify SNPs at appropriate distances, resequencing was carried out on 2 kb intervals spaced at 1, 5, 10, 20, 40, 80 and 160 kb in a single direction from each core SNP. Discovery was carried out in set of 44 unrelated individuals from Utah. The core SNPs and newly identified SNPs were then typed in Swedish and Nigerian samples.

Lander and colleagues [2] emphasize several important features of these data. Foremost is the considerable distance over which appreciable linkage disequilibrium occurs. Specifically, they show that a simple measure of linkage disequilibrium (D') retains an average absolute value above 0.5 between sites separated by up to 60 kb in Northern Europe. While the amount of linkage disequilibrium necessary for mapping would depend on the sample size and the relative risk of causal variants, a $|D'|$ of 0.5 would be a 'usable' [4] amount of linkage disequilibrium in some study designs. Interestingly, in the Nigerian sample, linkage disequilibrium was found to extend about an order of magnitude less far than in the Europeans, with the average $|D'|$ dropping below 0.5 at less than 5 kb.

As suggested by Reich *et al.* [2], the genomic extent of European linkage disequilibrium makes systematic and exhaustive gene mapping based on linkage disequilibrium a realistic prospect, as markers could be spaced every 60 kb, a roughly 20-fold reduction over theoretical predictions that assume no extreme demographic events in recent human history [4]. But extensive linkage disequilibrium is not strictly salutary; while it reduces the necessary scale of genome-wide screens, it can also make fine mapping more difficult. The limited linkage disequilibrium in Nigeria, however, is interesting in this regard. If such wide discrepancies in the extent of linkage disequilibrium among human populations holds up under further study, it should be possible to carry out coarse mapping in populations with more extensive linkage disequilibrium and fine mapping in populations with less extensive linkage disequilibrium [5], assuming that genetic causation is sufficiently similar across populations. The importance of populations with different characteristics extends to the use of populations with different patterns of linkage

disequilibrium to test whether associated variants are causal, as well as the use of putatively homogenous populations to reduce genetic heterogeneity and thereby increase the power of detection of variants with moderate effect.

Variance of linkage disequilibrium

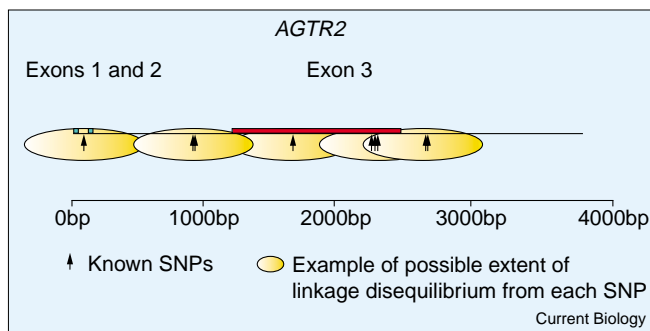
While it has attracted most attention in the literature, the average extent of linkage disequilibrium is not the only issue. More important is the variance of linkage disequilibrium as a function of distance between sites. Consider two different scenarios, both consistent with the average values across regions reported by Reich *et al.* [2]. In the low variance scenario, in a given region, all sites separated by less than say 10 kb are almost always in strong linkage disequilibrium, while sites farther than say 20 kb never are. In the high variance scenario, the full range of linkage disequilibrium values are observed in both distance classes, but with a greater tendency toward higher values at shorter distances.

How would this difference affect study design and interpretation? Under the low variance scenario, a single marker would 'mark off' a 10 kb interval, because it would be very likely to be in linkage disequilibrium with any variant within that interval. Perhaps even more importantly, under this scenario, when an association is observed it means that the causal variant cannot be farther away than 20 kb. Under the high variance scenario, however, neither of these conditions obtain. A single marker is insufficient to mark off the 10 kb interval because it cannot be counted upon to be in linkage disequilibrium with any given variant within the interval, and if an association is observed it is possible that the causal variant is a long way away. Thus the average extent of linkage disequilibrium is only the first step, and in fact the easiest. While we do not yet know which scenario is generally closer to reality, and whether it differs among populations, we have examples which appear closer to each [6,7].

The greater contribution of the Reich *et al.* [2] study therefore is what it can tell us about the variance of linkage disequilibrium. Because multiple sites in multiple regions are considered, these data will provide the best picture yet of the distribution of linkage disequilibrium as a function of distance in multiple genome regions, which will greatly facilitate the design of future studies. The data will contribute to the evaluation of factors influencing linkage disequilibrium, and can help assess how well patterns in genomic regions that are not yet studied can be predicted from those that have been studied. In fact, the importance of variation in recombination is already highlighted by the reported correlation with the extent of linkage disequilibrium [2]. Future studies will be required to assess other factors, such as gene proximity.

The importance of the distribution of linkage disequilibrium raises the question of power in association studies, which

Figure 1



Schematic representation of the *angiotensin receptor 2* (*AGTR2*) gene structure, showing locations of known SNPs. The ovals represent the possible extent of linkage disequilibrium from each known SNP.

currently lacks an appropriate framework for discussion. When an association study is carried out using an incomplete set of SNPs, we are interested in the extent to which the SNPs that are typed 'cover' those that are not through linkage disequilibrium. Knowledge of the distribution of linkage disequilibrium in particular genomic regions will provide the missing ingredient for assessing coverage. For example, Figure 1 shows the nine SNPs described in the database dbSNP (<http://www.ncbi.nlm.nih.gov/SNP>) that fall within the transcribed region of the *AGTR2* gene. The ovals represent an arbitrarily chosen average extent of linkage disequilibrium (in this case very short for the sake of illustration). In addition to these known SNPs, in any given population there is likely to be another set of SNPs that are not currently known. If we were to type only the known SNPs, we face the question of whether these would be sufficient to detect the presence of unknown SNPs with some specific relative risk in a given study design. As drawn here, we would have a chance of detecting SNPs within the transcribed region, but we might well miss something lurking in the regulatory region.

Regarding power of detection, under the low variance scenario all sites within an oval in Figure 1 might be in full linkage disequilibrium with the known SNPs. In this case a study using only the known SNPs in *AGTR2* would have exactly as much power of detection as would a study typing all the SNPs in this gene, which would therefore be less efficient. Under the high variance scenario, however, linkage disequilibrium might be much less consistent, even over such short intervals as indicated here. In this case, it might be necessary to have more than one known SNP within each oval, or it might be necessary to ensure that the ovals overlap, as they do in the third exon, to obtain the desired power. Clearly regions near multiple markers — where the ovals overlap — are better covered than those near only a single SNP.

For all of these reasons, a detailed knowledge of the variance of linkage disequilibrium will be essential in order to develop an appropriate statistical framework for interpreting association studies. Such a framework will not only aid study design and interpretation, but is also a prerequisite for being able to declare that a gene does not harbor an unknown variant conferring a relative risk greater than some specified level. Providing a systematic description of linkage disequilibrium will not be a trivial undertaking. The likelihood of variations across genomic regions due to differences in genealogical history, recombination rates or selection suggest that each region included in a genotype–phenotype association study should be assessed in its own right in the relevant population. In other words, what is ultimately needed is a genome project for linkage disequilibrium. But just as with the Human Genome Project, there are difficult questions about priority. Would it better to concentrate first on the small fraction of the genome represented by genes (accompanied by a generous upstream interval to capture regulatory regions)? Given the importance of cross-population comparisons, would it be better to gather less information from multiple populations, or to develop more complete pictures of fewer populations and build from there?

Most likely the community will pursue a hybrid strategy, for while we work out the optimal design for genome-wide work we also want guidance for more focused studies in the near term. For example, in certain cases relevant pathways are known in detail — as in the case of the renin–angiotensin pathway and variable efficacy of anti-hypertensive drugs — and genes within them are appropriate starting points for genotype–phenotype association studies. When exhaustive sets of SNPs finally do become available, one might wonder whether linkage disequilibrium can then be ignored. The answer is no. Even then, knowledge of the distribution of linkage disequilibrium will be required in order to translate detected associations between SNPs and phenotypes into quantitative statements about the possible genomic locations of the causal variant consistent with the observed associations.

Demography and linkage disequilibrium

Whatever the real patterns of linkage disequilibrium in human populations, they have been shaped by our evolutionary history. While linkage disequilibrium work is clearly focused on the development and refinement of mapping strategies, the data will also prove an unprecedented resource for evolutionary inference. For example, based on an idealized model of human history, Kruglyak [4] concluded that linkage disequilibrium would not extend much beyond about 3 kb, but also noted that this could be significantly greater in populations that have undergone a severe bottleneck. Recent admixture has also been shown to result in extended regions of linkage disequilibrium [5].

Reich *et al.* [2] used a simulation approach to ask why the patterns of linkage disequilibrium are so different in the North European and Nigerian samples. They found that admixture is unlikely to be the sole explanation, because the linkage disequilibrium observed in Europeans is greater than what admixture, even between distantly related human populations, is likely to produce.

Thus, it appears reasonable that a genetic bottleneck has affected the European but not African populations. The time of the bottleneck cannot be precisely pinned down. Reich *et al.* [2] suggest that it may be associated with founder events during the last glacial maximum in Europe, which leads to the testable prediction that Northern Europe has its own pattern of linkage disequilibrium, distinct from that of other non-African populations. Alternatively, the bottleneck could be associated with the emergence of modern humans from Africa, in which case the pattern of linkage disequilibrium it left would be shared among many different human populations [8]. Of course, multiple demographic events may have contributed to linkage disequilibrium in any given population and the relative contributions of such events could differ not only among populations, but throughout the genome. For example, the effect of a specific demographic event would last longer for tightly linked than for loosely linked pairs of sites. Assessing the demographic factors responsible for the linkage disequilibrium observed in Europe and elsewhere will require further in depth studies on other populations. Reich *et al.* [2] have provided an invaluable first step, but a comprehensive linkage disequilibrium map, of the whole genome and in multiple populations, is still a long way away.

Acknowledgements

We thank Alice Smith and Fiona Gratrix for technical assistance, and Neil Bradman and Michael Stumpf for comments on the manuscript.

References

1. Lazarou J, Pomeranz BH, Corey PN: Incidence of adverse drug reactions in hospitalised patients: a meta-analysis of prospective studies. *JAMA* 1998, **279**:1200-1205.
2. Reich DE, Cargill M, Bolk M, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian F, Ward R, Lander ES: Linkage disequilibrium in the human genome. *Nature* 2001, **411**:199-204.
3. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996, **273**:1516-1517.
4. Kruglyak L: Prospects for whole-genome linkage disequilibrium mapping common disease genes. *Nat Genet* 1999, **22**:139-144.
5. Wilson JF, Goldstein DB: Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am J Hum Genet* 2000, **67**:926-935.
6. 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 the human lipoprotein lipase. *Am J Hum Genet* 1998, **63**:595-612.
7. Rieder MJ, Taylor SL, Clark AG, Nickerson DA: Sequence variation in the human angiotensin converting enzyme. *Nat Genet* 1999, **22**:59-62.
8. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, BonneTamir B, Santachiara Benerecetti AS, Moral P, Krings M, *et al.*: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 1996, **217**:1380-1387.