Prospects and pitfalls in whole genome association studies

Robert W. Lawrence, David M. Evans and Lon R. Cardon*

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

Recent large-scale studies of common genetic variation throughout the human genome are making it feasible to conduct whole genome studies of genotype–phenotype associations. Such studies have the potential to uncover novel contributors to common complex traits and thus lead to insights into the aetiology of multifactorial phenotypes. Despite this promise, it is important to recognize that the availability of genetic markers and the ability to assay them at realistic cost does not guarantee success of this approach. There are a number of practical issues that require close attention, some forms of allelic architecture are not readily amenable to the association approach with even the most rigorous design, and doubtless new hurdles will emerge as the studies begin. Here we discuss the promise and current challenges of the whole genome approach, and raise some issues to consider in interpreting the results of the first whole genome studies.

Keywords: whole genome association; complex trait; genetic association

1. INTRODUCTION

Following the completion of the human genome sequencing project, a series of increasingly large studies have been conducted involving extensive genotyping of genetic variants around the genome (Patil *et al.* 2001; Dawson *et al.* 2002; Gabriel *et al.* 2002; Phillips *et al.* 2003; Ke *et al.* 2004; Hinds *et al.* 2005). The largest of these in the public domain is called the Human Haplotype Mapping Project (HapMap; Gibbs *et al.* 2003), an ongoing project that is genotyping samples from multiple populations on nearly all common genetic variants discovered to date. It has the potential to yield insights into patterns of intra- and inter-population variation in humans, recombination rates and hotspots, regions of possible historical selection and correlations of DNA sequence characteristics with the interesting locations of genetic variation. The main deliverable in the short-term, however, is simpler—a validated set of variant sites for subsequent disease-related investigations. By validating the genetic markers and identifying those that provide non-redundant information, the project should facilitate large-scale investigations of genotype–phenotype correlations. Such studies, known as whole genome association (WGA) studies, will involve screening hundreds of thousands of genetic markers on large samples of individuals having some phenotype of interest, e.g. a highly prevalent multifactorial disease or some continuous endo- or sub-phenotype(s).

Because of the availability of a very large set of common genetic markers, coupled with increases in genotyping capacity and decreasing genotyping costs (Chen & Sullivan 2003; Gunderson *et al.* 2005), there are a very large number of WGA projects currently

being planned, spanning a diverse array of human traits. There is a certain irony to this increasing enthusiasm, as genetic association studies are often viewed as having been largely unsuccessful in the past, despite 10 000 s of published reports (Weiss & Terwilliger 2000; Cardon & Bell 2001; Terwilliger *et al.* 2002). There is some debate in the field as to the actual ratio of ‘true positive’ to ‘false positive’ association reports (Lohmueller *et al.* 2003), but there have indeed been a very large number of unreplicated association studies (Ioannidis *et al.* 2001; Ioannidis *et al.* 2003). The difficulty in labelling such reports as unequivocal false positives comes from the fact that in genetic studies, there are many reasons for lack of reproducibility apart from type I error. Individual differences in mutations within a gene (allelic heterogeneity), between genes (genetic heterogeneity), between samples (population heterogeneity) and within phenotypes (phenotypic heterogeneity) can all lead to differences in study outcomes, so determining what is a true versus false association result remains a challenge and is an area of active research. Although there remain a number of hurdles for WGA studies, it is clear that they are imminent. Here we describe some of the recent results leading up to WGA studies, and offer some perspectives on the challenges to come.

2. LINKAGE DISEQUILIBRIUM

From a genomic perspective, part of the challenge in obtaining consistent association results stems from the considerable genetic diversity within populations. Differences in allele frequencies within and between groups have long been observed (Goddard *et al.* 2000; Shifman *et al.* 2003), and they have led to the highly publicized inferences of ancestral human migration patterns (Cavalli-Sforza *et al.* 1994) that have become popular science. Until recently, however, there was relatively little empirical information about the

* Author for correspondence (lon.cardon@well.ox.ac.uk).

One contribution of 12 to a Discussion Meeting Issue ‘Genetic variation and human health’.

structure of genetic variability; i.e. the patterns of variation throughout the genome. New mutations are co-transmitted with the pre-existing variants on their background chromosomes until they are broken apart by some form of chromosomal rearrangement, typically recombination (Petes 2001). Over generations, these processes of recombination yield mosaics of allelic variants, some of which have remained closely associated with each other over time, and others whose initial co-occurrence has been eroded, often completely.

Linkage disequilibrium (LD) refers to the degree to which two allelic variants are associated in a population. For markers A and B, it is usually summarized on the basis of the simple expression $D = P_{AB} - P_A P_B$, most commonly as $D' = D / \max(D)$ or $r^2 = D^2 / [P_A(1 - P_A)P_B(1 - P_B)]$ (Weir 1996). In general, there is a positive correlation between physical distance and LD, such that variants that are close together are more strongly correlated with each other than those with greater separation, as the former are less likely to be broken apart by recombination. However, one of the main outcomes of the recent large-scale genotyping projects, including the HapMap, has been to highlight the striking degree of variability in LD.

Figure 1 shows a plot of pairwise correlations between markers as a function of their physical separation. Overlaid on the scatterplot is a fitted curve showing the exponential decay of LD according to physical distance. The dramatic variability in LD indicates that despite the physical proximity of markers, knowing the allele status of one genetic variant often carries little information about neighbouring variants. Thus, it is inappropriate to consider 'genomic coverage' of genetic markers purely in terms of physical spacing, a fact which remains misunderstood by many disease investigators. This is relevant to disease association mapping because it implies that, in cases of low LD, one might identify the correct chromosome, gene and even intron, exon or regulatory element harbouring trait-influencing variants, but unless the exact aetiological variant is assayed, negative or unreproducible results may emerge. This is a daunting challenge in terms of genomic complexity, as human populations are thought to carry approximately 10 million common variants, and many more rare variants (Ardlie *et al.* 2002; Carlson *et al.* 2003, 2004). This variability has led to some suggestions that large-scale studies of genetic variation in one population will not be relevant for others, and that studies such as the HapMap may thus have limited utility for association studies (Sawyer *et al.* 2005).

3. HAPLOTYPE BLOCKS

In the past five years, several studies have put forth encouraging views regarding the complexities of LD (Daly *et al.* 2001; Jeffreys *et al.* 2001; Patil *et al.* 2001; Gabriel *et al.* 2002). As it became possible to examine markers closer together on chromosomes, consistent trends emerged that appeared less stochastic than the high degree of 'noise' in the pairwise LD/physical spacing relationship. Daly *et al.* (2001) described staccato patterns of alternating segments of DNA,

Figure 1. Variability and decay in linkage disequilibrium (LD) by physical spacing on chromosome 22. Pairwise D' coefficients are plotted against the physical separation of the markers, revealing a general trend of decay with distance, but also extensive variability. The curve overlaid on the scatterplot is from a fitted model of expected decay $E(D_{ij}) = D_{\text{low}} + (D_{\text{high}} - D_{\text{low}})(1 - r)^t$, where D_{low} , D_{high} and t (the number of generations of decay) are estimated from the data. The raw data and this model were described by Dawson *et al.* (2002).

with some regions exhibiting little haplotype diversity, punctuated by segments of very high diversity. In terms of LD, the regions of low diversity, which they termed 'haplotype blocks', were generally high, (presumably) owing to less recombination than the intervening regions of low LD.

The haplotype block concept has rapidly gained popularity (Paabo 2003), as it offers an attractive intuitive framework for disease-gene mapping (Goldstein 2001; Gabriel *et al.* 2002). In theory, if one could identify the haplotype blocks across the genome, then disease-association studies could use customized sets of genetic markers to capture the information within and between them. However, increasingly dense datasets have suggested that, while the haplotype block model provides a useful view of broad patterns of LD, specific attributes of blocks, particularly block boundaries, can be strongly influenced by the markers and human samples chosen for analysis (Phillips *et al.* 2003; Wall & Pritchard 2003; Ke *et al.* 2004). There are also multiple statistical definitions of blocks which yield similar, though non-identical patterns (Schulze *et al.* 2004; Zeggini *et al.* 2005; Ding *et al.* 2005). Thus, for specific association-design/interpretation questions, such as the extent to which one might evaluate association across a candidate gene, the variability in boundaries and definitions can induce further uncertainty into association studies. Haplotype blocks provide a useful platform for considering patterns of genetic variation, but they are not a panacea for WGA studies.

4. RECOMBINATION RATES

Although the block model may not be especially helpful for focused marker questions, it does highlight non-uniform patterns of LD in the human genome. Presumably, block patterns are largely generated from ancestral recombination, though pairwise LD measures are suboptimal in this regard as they confound recombination history with a number of factors influencing the frequency of the alleles, such as drift, selection, non-panmictic mating, etc. (Zavattari *et al.* 2000; Zhang *et al.* 2004). Accordingly, there has been

Figure 2. Comparison of population recombination rates and pairwise LD on chromosome 20. The left column shows estimates of recombination rates between adjacent markers and the right column shows D' coefficients for the same markers. Panel (a) plots values for Asian samples (y -axis) against Caucasians (x -axis). Panel (b) plots values for African Americans against Caucasians. Panel (c) plots values for African Americans against Asians. The samples are described in (Ke *et al.* 2004). The population recombination rates are described in detail in (Evans & Cardon 2005).

considerable interest in recent developments of statistical methods that directly estimate population recombination rates from unphased genotyped data (Fearnhead & Donnelly 2001; McVean *et al.* 2002; Clark *et al.* 2003; Li & Stephens 2003). In comparison with LD measures, population recombination rates show greater consistency both within and between populations. This consistency is shown in figure 2, using our data from a recent study of chromosome 20 (Evans & Cardon 2005). Pairwise LD varies substantially between populations, mainly because of allele frequency differences amongst populations. This is particularly evident with the D' measure. In contrast, while average levels of recombination vary in different groups, regions of high (low) recombination tend to be remarkably similar. Other studies have demonstrated similar trends (e.g. Clark *et al.* 2003; Crawford *et al.* 2004; McVean *et al.* 2004).

Empirical studies of recombination hotspots have suggested that they often take place across relatively short segments of DNA (Jeffreys *et al.* 2000; Jeffreys *et al.* 2001). These fine-scale changes, coupled with the

apparent similarity of high/low regions across human samples (Ke *et al.* 2004), suggest that recombination patterns may provide useful information for allelic association studies. Curiously, preliminary work on the locations of recombination hotspots suggest that they are not particularly well aligned between humans and other primates, raising intriguing questions about natural selection and human evolution (Ptak *et al.* 2005; Winckler *et al.* 2005). Also, there is some evidence that hotspots can occur even within haplotype blocks or regions of otherwise high LD (Jeffreys *et al.* 2005). Overall, the recombination studies are revealing exciting and unexpected trends. How this information can inform or be integrated into association studies is presently a topic of concerted interest.

5. WHOLE GENOME ASSOCIATION STUDIES

On balance, while there are a number of stimulating results emerging from the HapMap and other large-scale genotyping studies, there remain many challenges for WGA studies (Hirschhorn & Daly 2005; Wang *et al.* 2005). At present, we are not in a position to clearly

Figure 3. Increase in availability of genetic markers. The number of markers available in dbSNP is shown as a function of the dbSNP release date (all data from www.ncbi.nlm.nih.gov/SNP). The full distribution (dark grey; greater than 10 million in 2005) reflects all non-redundant SNPs in dbSNP, while the distribution in light grey shows the number of validated SNPs.

delineate precise genomic regions that are worthy of greater (lesser) attention in disease-gene mapping, nor are we able to safely depict the boundaries of association or linkage intervals. The output of most immediate and practical relevance for association studies is a vast set of genetically validated markers (figure 3). This is an impressive achievement, considering that only a few years ago, candidate gene studies required the use of restriction fragment length polymorphisms (RFLPs) or extensive resequencing to find even common-allele markers. The single nucleotide polymorphism (SNP) information is valuable to nearly all association studies, family- or population-based, and involving candidate genes or the entire genome.

Unfortunately, while the validated markers will make WGA studies cheaper, faster and more accessible, they do not guarantee greater success. Small sample sizes have long plagued allelic association studies, and genotyping more markers does not compensate for a lack of power to detect real effects in the first place. When the expected effect size of each locus is small, as in common complex traits, large numbers of individuals are required whatever the density of markers genotyped (Risch & Merikangas 1996; Zondervan & Cardon 2004). This sampling problem is exacerbated in the context of WGA, as the study of many markers creates multiple-testing problems with the correlated variants. WGA studies generally require larger, not smaller, sample sizes than previous investigations of specific candidate genes or regions.

Nevertheless, arguments for or against WGA studies, or even descriptions of the likely effects of genetic variability and sample sizes noted above, are mainly theoretical as, to date, there have been only a few association studies involving extensive marker coverage (Ozaki *et al.* 2002; Roses *et al.* 2005), and none that have spanned the entire genome (i.e. coding plus non-coding regions). Without real data, it is difficult to design and test new methods of analysis, compare strategies and study designs and consider the diversity of genetic architectures for different phenotypes.

To gain an initial view of what a WGA study might look like, we merged two publicly available datasets: one involving gene expression (Morley *et al.* 2004) and

another involving interim data from the HapMap project (Gibbs *et al.* 2003). We considered the 100 most heritable gene-expression values as phenotypes (Y), which we regressed on each diallelic HapMap marker one-at-a-time: $E(Y) = a + bG_j$, where G_j was coded to detect additive genetic effects (0, 1, 2 for genotypes aa, Aa, AA, respectively) for the j th marker. Our analyses are not meant to represent a true genome-wide study nor a thorough assessment of the factors influencing gene expression, as the association sample size (42 unrelated individuals) is too small for such consideration and the HapMap data are not yet complete for a full assessment. Rather, we are interested in simply gaining an initial view of any trends that might emerge in a study of WGA.

The results of the pseudo-WGA studies revealed a number of encouraging findings. For most of the expression phenotypes that showed evidence for family-based linkage, evidence for strong association (considered as $p < 10^{-9}$ due to conservative Bonferroni correction) was observed in at least one location in the genome. However, this was not always the case, even with a very high density of markers. In some cases, despite strong evidence for linkage and association testing of more than 500 000 non-redundant SNPs, no evidence for allelic association was apparent. In this case, further genotyping, resequencing or a non-association based strategy (e.g. for rare alleles) would be required to identify the variants contributing to the linkage profiles, as the WGA approach would fail with the available data.

6. GENETICALLY IDENTICAL SNPs

Another interesting pattern emerged that has potential implications for gene-localization in WGA studies (figure 4). In this example, which was highlighted as a case of *cis*-linkage in the initial expression study (Morley *et al.* 2004) since the primary linkage evidence was obtained in the same chromosomal region of the gene itself, several loci are apparent in association analyses (figure 4a). Interestingly, none of these are located within the linkage region, i.e. any *cis*-acting loci leading to the linkage signal are not apparent, despite strong association evidence for several other loci in the genome. This may reflect the well-rehearsed difficulties that can accompany attempts to refine bona fide

Figure 4. Initial profile of whole genome association results: Effects of giSNPs. (a) An eQTL (gene expression level as a quantitative trait locus) association scan using only non-redundant markers (i.e. those for which $r^2 < 1.0$). (b) The same results as (a) but including the redundant markers. The peaks of identical amplitude in (b) reflect the genotypic identity between the markers. From the data at hand, it is not clear which of the peaks, if any, reflect the aetiological alleles, thus emphasizing the difficulties with location inference in association studies. The expression data used in these analyses are described in Morley *et al.* (2004); the HapMap genotype data are from the December 2004 release (www.hapmap.org).

linkage regions by population-based association analysis (Terwilliger *et al.* 2002).

Perhaps more interesting is the role of LD in the patterns observed. In figure 4a, the analyses were conducted only on non-redundant HapMap markers; i.e. whenever two or more markers were identical (every individual had the same genotype at marker B as they had at marker A, or $r^2 = 1.0$), the first one observed was arbitrarily retained and the others omitted as they give identical association information. With this marker selection, the association analyses showed strong evidence on chromosomes 9, 12 and 17. In general, the redundant markers, or 'genetically identical SNPs' (giSNPs), are located in close proximity owing to the general positive correlation between LD and physical spacing (Lawrence *et al.* submitted). However, as noted above, there is a high degree of variability in this relationship.

The impact of the LD variability on association results are exemplified in figure 4b, as there are giSNPs located several megabases apart. In this particular case, the most strongly associated SNPs are also giSNPs, yielding identical association evidence at distant parts of each chromosome. Which location is correct? It is not possible to discern this from the association evidence alone, which raises a potentially serious concern about inferring the location of aetiological variants from association evidence. One could resequence all variants around specific genes and still miss the active site since it is located far away. This situation resembles that of point mutations in long-range controlling elements, which have also been observed at megabase distances from their target (Lettice *et al.* 2002). In the case of LD, the increasing density of the HapMap and other projects are indicating that more than 50% of common-allele SNPs have at least one giSNP partner (Lawrence *et al.* submitted; Lettice *et al.* 2003). Thus, it is possible that this particular example may not be unusual. It will be important to maintain awareness of such eventualities in the imminent WGA studies.

7. CONCLUSIONS

Substantial advances in genotyping and international efforts to validate common genetic variants have led to renewed interest in association studies. Accordingly, whole genome scans for human complex trait loci are now feasible and imminent. This is an exciting opportunity for human genetics, as it may generate a number of new insights about the nature of individual differences. It is important to recognize, however, that such studies are not guaranteed to identify new genes. Practical issues such as distinguishing true from false claims of replication, small sample sizes, long-range marker correlations and difficulties with robust inference based on large numbers of statistical tests pose real challenges for WGA studies. Moreover, there are a number of situations in which association studies, however well designed, are poorly suited for identification of aetiological variants; e.g. when the causal alleles are rare and heterogeneous. On balance, whole genome association studies seem highly likely to yield novel findings. The extent to which these findings are robust and lead to new disease insights will depend at least in part in our ability to manage the challenges at hand and remain open to those that are certain to emerge.

This work was supported by the Wellcome Trust and the US National Institutes of Health. We thank Professor Bruce Weir and Dr Dahlia Nielsen for collaboration on the expression analyses and the HapMap analysis group for helpful discussions about LD and many other characteristics of the genome. We also thank Drs David Bentley and Panos Deloukas at the Wellcome Trust Sanger Institute for many fruitful collaborations contributing to this research.

REFERENCES

- Ardlie, K. G., Kruglyak, L. & Seielstad, M. 2002 Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309.
- Cardon, L. R. & Bell, J. I. 2001 Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99.
- Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L. & Nickerson, D. A. 2003 Additional SNPs

- and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521.
- Carlson, C. S., Eberle, M. A., Kruglyak, L. & Nickerson, D. A. 2004 Mapping complex disease loci in whole-genome association studies. *Nature* **429**, 446–452.
- Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. 1994 *History and geography of human genes*. Princeton: Princeton University Press.
- Chen, X. & Sullivan, P. F. 2003 Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Pharmacogenomics J.* **3**, 77–96.
- Clark, A. G., Nielsen, R., Signorovitch, J., Matise, T. C., Glanowski, S., Heil, J., Winn-Deen, E. S., Holden, A. L. & Lai, E. 2003 Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.* **73**, 285–300.
- Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A. & Stephens, M. 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genet.* **36**, 700–706.
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. 2001 High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232.
- Dawson, E. *et al.* 2002 A first generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548.
- Ding, K., Zhou, K., Zhang, J., Knight, J., Zhang, X. & Shen, Y. 2005 The effect of haplotype-block definitions on inference of haplotype-block structure and htSNPs selection. *Mol. Biol. Evol.* **22**, 148–159.
- Evans, D. M. & Cardon, L. R. 2005 A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet.* **76**, 681–687.
- Fearnhead, P. & Donnelly, P. 2001 Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.
- Gabriel, S. B. *et al.* 2002 The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.
- Gibbs, R. A. *et al.* 2003 The International HapMap Project. *Nature* **426**, 789–796.
- Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. 2000 Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234.
- Goldstein, D. B. 2001 Islands of linkage disequilibrium. *Nature Genet.* **29**, 109–111.
- Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G. & Chee, M. S. 2005 A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genet.* **37**, 549–554.
- Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. 2005 Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079.
- Hirschhorn, J. N. & Daly, M. J. 2005 Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108.
- Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. 2001 Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309.
- Ioannidis, J. P., Trikalinos, T. A., Ntzani, E. E. & Contopoulos-Ioannidis, D. G. 2003 Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**, 567–571.
- Jeffreys, A. J., Ritchie, A. & Neumann, R. 2000 High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733.
- Jeffreys, A. J., Kauppi, L. & Neumann, R. 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222.
- Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S. & Donnelly, P. 2005 Human recombination hot spots hidden in regions of strong marker association. *Nature Genet.* **37**, 601–606.
- Ke, X. *et al.* 2004 The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**, 577–588.
- Lawrence, R. W. *et al.* Submitted. The influence of genetically identical SNPs on inferring the location of disease-associated variants.
- Lettice, L. A. *et al.* 2002 Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548–7553.
- Lettice, L. A. *et al.* 2003 A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735.
- Li, N. & Stephens, M. 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.
- Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182.
- McVean, G., Awadalla, P. & Fearnhead, P. 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.
- McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.
- Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. & Cheung, V. G. 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747.
- Ozaki, K. *et al.* 2002 Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654.
- Paabo, S. 2003 The mosaic that is our genome. *Nature* **421**, 409–412.
- Patil, N. *et al.* 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723.
- Petes, T. D. 2001 Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2**, 360–369.
- Phillips, M. S. *et al.* 2003 Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genet.* **33**, 382–387.
- Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A. & Paabo, S. 2005 Fine-scale recombination patterns differ between chimpanzees and humans. *Nature Genet.* **37**, 429–434.
- Risch, N. & Merikangas, K. 1996 The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.
- Roses, A. D., Burns, D. K., Chisoe, S., Middleton, L. & St Jean, P. 2005 Disease-specific target selection: a critical first step down the right road. *Drug Discov. Today* **10**, 177–189.

- Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J. & Kidd, K. K. 2005 Linkage disequilibrium patterns vary substantially among populations. *Eur. J. Hum. Genet.* **13**, 677–686.
- Schulze, T. G., Zhang, K., Chen, Y. S., Akula, N., Sun, F. & McMahon, F. J. 2004 Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum. Mol. Genet.* **13**, 335–342.
- Shifman, S., Kuypers, J., Kokoris, M., Yakir, B. & Darvasi, A. 2003 Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* **12**, 771–776.
- Terwilliger, J. D., Haghghi, F., Hiekkalinna, T. S. & Goring, H. H. 2002 A biased assessment of the use of SNPs in human complex traits. *Curr. Opin. Genet. Dev.* **12**, 726–734.
- Wall, J. D. & Pritchard, J. K. 2003 Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73**, 502–515.
- Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. 2005 Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118.
- Weir, B. S. 1996 *Genetic data analysis II*. Sunderland, MA: Sinauer Associates.
- Weiss, K. M. & Terwilliger, J. D. 2000 How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157.
- Winckler, W. *et al.* 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111.
- Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J. A. & Cucca, F. 2000 Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Hum. Mol. Genet.* **9**, 2947–2957.
- Zeggini, E., Barton, A., Eyre, S., Ward, D., Ollier, W., Worthington, J. & John, S. 2005 Characterisation of the genomic architecture of human chromosome 17q and evaluation of different methods for haplotype block definition. *BMC Genet.* **6**, 21.
- Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., Bentley, D. R. & Morton, N. E. 2004 Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl Acad. Sci. USA* **101**, 18 075–18 080.
- Zondervan, K. T. & Cardon, L. R. 2004 The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5**, 89–100.