Commentary

New hope for haplotype mapping

John I Bell

John Radcliffe Hospital, Oxford, UK

Corresponding author: John Bell (e-mail: Regius@medsci.ox.ac.uk)

Received: 1 November 2002 Accepted: 20 November 2002 Published: 13 January 2003

Arthritis Res Ther 2003, **5**:51-53 (DOI 10.1186/ar621)

Abstract

The systematic analysis of polymorphisms across large parts of the human genome has begun to provide the first information on haplotypes and the problem of linkage disequilibrium across large genomic regions. These data suggest that significant regions of the gnome show highly conserved haplotypes, potentially enhancing the ability to detect disease associations.

Keywords: evolution, genetics, haplotypes, human leukocyte antigen

Introduction

Individual risk of developing most major diseases can be largely attributed to the extensive single nucleotide variation that occurs throughout the human genome. The identification of the functional variants that contribute to disease risk and progression, however, has been difficult, particularly for complex diseases where the interplay of genes and environment is most evident.

Relatively minor degrees of genetic variation can lead to substantial structural and functional changes – as evidenced by the modest changes that distinguish primate species or that can produce profound disease phenotypes in Mendelian-related traits. Attempts to identify DNA variants that contribute to complex disease through linkage analysis with genome wide markers in families have provided localisation of large genetic effects, but few actual disease-mediating polymorphisms. Association strategies, including genome wide association, provide a theoretically more powerful methodology for identifying disease polymorphisms, but also present new methodological and statistical challenges. These have, however, provided hope that such variants can now be identified.

One challenge in applying association methodology is to identify functional variants without analysing every polymorphism in a genomic region, which may be as frequent as 1/1000 base pairs in regions of the genome. If all the polymorphisms had achieved equilibrium through recombination with each other, so that adjacent polymorphisms occur together at a frequency determined only by their allele frequency, this task would be enormous. Fortunately, for much of the genome the distribution of alleles is not in equilibrium, reducing the scale of the challenge of extracting all the necessary genetic information from some genomic regions.

The occurrence of a set of polymorphisms along a single chromosome is referred to as a haplotype. The frequency with which polymorphisms reside together on a haplotype is dependent on a number of factors: the evolutionary history of the population studied, the recombination frequency and recombination hot-spots sites along the chromosome, and the evolutionary selection of advantageous or disadvantageous functional variants. When alleles at adjacent sites are found together more often than would be expected if the region were in equilibrium, they are said to be in linkage disequilibrium (LD). The result of LD is that particular combinations of alleles are conserved across haplotypes, and typing any one of these will provide information across the whole haplotype. The obvious benefit is that information about association can be attained across large genomic regions by typing only very small numbers of single nucleotide polymorphisms.

The importance of LD for those interested in finding disease genes in the genome is well illustrated by the human leukocyte antigen (HLA) region. Genetic typing

HLA = human leukocyte antigen; LD = linkage disequilibrium.

was available here long before molecular genetic technologies arrived because the polymorphism on these genes was recognisable through the use of serological reagents. Early studies revealed the association between individual alleles and human disease. For example, the earliest associations between HLA and type I diabetes revealed that HLA B8 was associated with the disease. As typing became widespread, it became clear that the HLA region on chromosome 6 was a genomic region that contained strong LD. This meant that certain alleles could define ancestral haplotypes with LD extending over very large distances (up to 3 cM) and that the association of any one of many alleles could implicate a haplotype associated with disease. This led to the rapid association of the A1 B8 DR3 haplotype with a range of autoimmune disorders, including diabetes in Caucasian populations. Eventually, the true functional variants that confer susceptibility to type I diabetes were shown to arise from the HLA class II region, a megabase away from the those variants originally shown to associate with disease. Most other HLA disease associations relied upon LD initially to be identified. Thirty years later, these associations remain the best examples of complex trait genetic associations to be documented, despite years of molecular genetic mapping.

It has been assumed by many that the extent of LD surrounding the HLA was special and that the lessons learned from exploring the disease gene in this region of high allelic association would not be applicable to the rest of the genome. As attention in disease gene hunting moved from genome wide linkage studies to the exploration of linked regions, and as the idea of whole genome association as a plausible method for identifying disease polymorphisms arose, there has been renewed interest in establishing how much LD exists elsewhere in the genome. If there were extensive regions outside the HLA that could be defined by a relatively small number of markers, the job of identifying regions containing disease genes would be made much easier. Large regions of the genome could then be scanned with existing technology, without it being necessary to type every DNA variant individually in an attempt to identify the functional polymorphism responsible for a disease.

Until recently, only a few studies provided limited information about the extent of LD around the genome. Two publications have appeared that provide an indication of LD; one having typed DNA variants in 51 autosomal regions of the genome [1], and the other having intensively typed polymorphisms across the whole of the long arm of chromosome 22 [2]. These two publications provide our first glimpses into the haplotypes that might exist within the genome and have important implications for our ability to map disease genes in the near future. Interestingly, these publications have taken rather different approaches to their studies and have generated somewhat different conclusions.

Gabriel *et al.* [1] analysed 3738 polymorphisms in a range of ethnic groups across 51 autosomal regions averaging 250,000 base pairs in length. Their paper identified many haplotype blocks, defined as a region over which a very small proportion (<5%) of comparisons among informative single nucleotide polymorphisms show strong evidence of historical recombination. This is an extremely rigorous test of LD, requiring almost complete allelic association across the haplotypes. Gabriel *et al.* used markers at close intervals (on average every 7.8 kb) and, as a result, generated data on a large amount of LD that is known to occur at short intervals. The vast majority of the haplotype blocks defined in this study were in regions <5 kb, a distance well recognised to be associated with strong LD in Caucasian populations. The extreme criteria for defining haplotypes contributed to Gabriel *et al.*'s observation that LD does appear to decline with the distance between markers within a haplotype block. This study is largely measuring almost pure, conserved haplotypes that, on average, are 11 kb in length in Nigerian and Afro-American samples, and are 22 kb in length in European and Asian samples. These haplotypes could be identified by as few as six to eight markers. Based on these data, the authors estimate that 300,000–1,000,000 single nucleotide polymorphisms would be necessary to have a fully powered genome wide association strategy using this sort of haplotypic information.

Dawson *et al.* [2] took a different approach that results in significantly different conclusions. They used markers that, on average, are 15 kb apart across the whole of the long arm of chromosome 22. This study was able to look at much larger regions of LD, using 1504 markers across the chromosome and using conventional measures of LD (D' and r^2) rather than the more stringent criteria used by Gabriel *et al.* [1]. This provides evidence for haplotype blocks that are less pure, but extend over much longer regions. As one would expect, LD decays over increasing distance in these haplotypes. The regions of extensive LD correlate with regions of the chromosome known to have low recombination rates. The longest haplotype network seen by this group was 804 kb in length containing 16 markers, while 25 markers make up a haplotype network of 758 kb elsewhere on the chromosome. These are not completely pure haplotypes, but represent regions where low rates of recombination have, in European populations, long conserved haplotype networks that can be defined by a relatively small number of markers.

What then should the gene mappers conclude from these apparently disparate results? By defining haplotypes very rigorously, one will find many short stretches of virtually complete LD in the genome. A less stringent approach can establish the presence of longer ancestral haplotypes across which the levels of LD vary, but which reduce the complexity of genotyping necessary to describe the region. The best way to evaluate what might be valuable is to again review what has already proved useful in the HLA.

Although it has not been demonstrated that the LD across the HLA is broken up by punctate regions of recombination, the haplotypes and LD patterns that have helped define disease associations often operate across these sites. Long-range LD has proved powerful as many class II associations originated with class I associations. None of these HLA haplotypes are complete or pure; most represent ancestral haplotypes on which new variants have arisen. In some cases, they extend from well beyond the HLA-A locus at one end to the HLA-DP at the other. Despite their size, they have proved immensely valuable in disease gene mapping. One would argue, therefore, that the approach used by Dawson *et al*. [2] may provide better estimates of what will be useful in real studies of disease genes.

It is important also to remember that, although LD and conserved haplotypes may assist in identifying regions associated with disease, it also makes the final identification of disease mutations more difficult. Regions of LD contain multiple DNA variants, all of which may be strangely associated with a disease, due to being on the same conserved haplotype. This can make the precise identification of the functional variant extremely difficult, as has been seen within the HLA. Only transracial studies that break down LD and conserved haplotypes can resolve these challenging issues.

Conclusion

Identifying disease-related genetic polymorphisms in common disease has never been easy. Recognising, however, that patterns of LD that were previously thought confined to the HLA are in fact much more widespread should greatly facilitate the introduction of hypothesis-free association strategies.

Competing interests

None declared.

References

1. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyomo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225-2229
2. Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lõhmussaar E, Zernant J, Tõnisson N, Remm M, Mãgl R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I: **A first-generation linkage disequilibrium map of human chromosome 22.** *Nature* 2002, **418**:544-548.

Correspondence

John I Bell, Regius Professor of Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK. Tel: +44 1865 221340; fax: +44 1865 220993; e-mail: Regius@medsci.ox.ac.uk