Supplement Review

Single nucleotide polymorphisms and disease gene mapping

John I Bell

Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford, UK

Correspondence: John I Bell, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK. Tel: +44 (0)1865 221340; fax: +44 (0)1865 220993; e-mail: john.bell@ndm.ox.ac.uk

Received: 27 February 2002
Accepted: 3 March 2002
Published: 9 May 2002

Arthritis Res 2002, **4 (suppl 3)**:S273-S278

(Print ISSN 1465-9905; Online ISSN 1465-9913)

Chapter summary

Single nucleotide polymorphisms are the most important and basic form of variation in the genome, and they are responsible for genetic effects that produce susceptibility to most autoimmune diseases. The rapid development of databases containing very large numbers of single nucleotide polymorphisms, and the characterization of haplotypes and patterns of linkage disequilibrium throughout the genome, provide a unique opportunity to advance association strategies in common disease rapidly over the next few years. Only the careful use of these strategies and a clear understanding of their statistical limits will allow novel genetic determinants for many of the common autoimmune diseases to be determined.

Keywords: disease association, genetics, HLA, linkage disequilibrium, SNP

Introduction

Advances in human molecular genetics have greatly enhanced our ability to identify the genetic basis for many human diseases, including the autoimmune diseases. Characterization of the genetic determinants requires the evaluation of polymorphism in families and populations to detect the relationship between individual variants and disease phenotypes. Such activities originally relied on measuring polymorphism at the protein level. The extensive literature on human leukocyte antigen (HLA) and disease began with antibodies that defined allelic variation with sera that recognized individual polymorphisms, now understood to be expressed in a host of HLA molecules that regulate the immune response.

This approach has recently become more general as DNA variants can be characterized systematically and can be correlated with disease. The first sets of polymorphisms available for such studies were variable repeats spread throughout the genome. Typing polymorphisms in DNA has always been a limiting factor in such studies and, for the first time, the identification of variable number of tandem repeat sequences allowed highly variable repeat sequences to be readily typed using Southern blots [1]. However, the infrequency of such complex repeats limited their utility and, although they have been informative in terms of describing individual genetic diversity, their role in identifying disease-related genetic variation has been limited.

The next revolution in genetic variation scoring arose directly from the introduction of another methodology; the polymerase chain reaction. This methodology allowed the systematic characterization of smaller dinucleotide and trinucleotide repeats throughout the genome. These could be easily identified, were highly polymorphic, and could be readily mapped at high density throughout the human genome. These repeats provided the first mechanism for searching genome wide for evidence of genetic linkage and disease. Their frequency was not sufficiently high, however, to allow their application in population-based studies of human disease that had proved so productive

when the HLA region had been studied using antibodies. Despite this, the application of these markers became the basis for whole genome linkage studies that have provided information, in family samples, of the localization of a large number of disease genes in autoimmune disease and other complex traits [2].

The most significant development to date in the molecular study of disease genetics has emerged from the availability of the human genome sequence. This data provided the template from which to generate extensive amounts of information on single nucleotide variants. Several important advantages emerge from the availability of such single nucleotide polymorphisms (SNPs). These are by far the commonest form of polymorphism within the genome. These variants will account for the vast majority of polymorphism responsible for human disease. The variation occurs in both coding and noncoding sequences at a frequency of approximately 1 per 1000 base pairs. The extent of variation is limited, however, by the complex relationship between these variants, reflecting population history, recombination hot spots, and selection.

Although large numbers of such polymorphisms had been previously described, both within genes and in intervening sequences, the systematic generation of very large numbers of such variants at high density throughout the genome was necessary if these variants were to be used to look systematically for disease-related genetic variation. The generation of this large set of SNPs has been accompanied by methodology for typing such variation efficiently. The binary nature of these polymorphisms makes them much easier to type in an automated fashion. As this methodology becomes more inexpensive and efficient, it should be possible to catalogue the polymorphisms that exist at a very large number of sites in individuals with and without disease, and hence should be possible to derive information about the genetic susceptibility factors that contribute to many human diseases.

This review will consider the role of these SNPs in mapping human disease related to DNA variation; in particular, the use of SNP typing to characterize haplotypes throughout the genome, and the use of linkage disequilibrium (LD) patterns to find changes that relate to phenotypes. It will also consider how such information should be used to study disease in association strategies and will highlight the basis for current failings of this approach.

The SNP map of the human genome

Although there is extensive data to suggest that single nucleotide variation exists within the human genome at a frequency of approximately 1 per 1000 base pairs [3], only recently has extensive sequencing and resequencing of the genome provided large sets of SNPs to study [4–6]. The publication reporting the identification of 1.42 million SNPs in 2001 provided the necessary information to study their relationship to each other across the genome [4].

The SNP Consortium (TSC) has contributed the largest set of SNPs identified by shotgun sequencing of genomic fragments [4]. These were obtained using an ethnically diverse panel of 24 individuals. Other large sets of SNPs have been derived from analysis of overlap regions of the human genome sequence derived from bacterial artificial chromosome-derived and P1-derived artificial chromosome sequences obtained during the human genome project. The allele frequency of these SNPs suggests that 82% are polymorphic at frequencies of 10% or greater in at least one ethnic population. Twenty-seven percent had an allele frequency $>20\%$ in the three ethnic groups studied (European, African-American and Chinese).

On average, the TSC SNPs were most evenly distributed and were found every 3.05 kb. The SNP density was found to be relatively even across most chromosomes except for the X and Y chromosomes, where there were lower rates of heterozygosity. This was explained by the lower effective population size associated with the X chromosome (hemizygosity in males means that the effective population is three-quarters that of the autosomes) and with the Y chromosome, which has a lower effective population size but a higher mutation rate at male meiosis. The pseudoautosomal region of the Y chromosome known to recombine with the X chromosome shows very high levels of heterozygosity as expected, while the nonrecombining region shows very low levels.

The extent of polymorphism is potentially driven by a number of factors. The population history would undoubtedly be important, in particular the timing of individual mutations, the population admixture and recombination rates, as will balancing selection in some cases. Regions of very high heterozygosity were also detected, including the HLA region on chromosome 6. We already know a great deal about this region and the role of balancing selection in establishing and maintaining polymorphisms in the HLA loci. It will be interesting to see whether other regions of high heterozygosity relate to genomic regions where selection and or recombination rates have played a dominant role.

LD patterns and haplotypes

With the availability of genome wide sets of SNPs, it becomes possible to search the genome for genetic variation that accounts for human disease and other traits [7,8]. No technology is currently available to type the full set of SNPs in population samples, nor indeed is such a comprehensive set of markers available. Of the marker sets available of polymorphic sequences, only a fraction of functional disease-related SNPs are currently available.

If each genetic polymorphism added independently to the variation, then the total set of variable chromosomes would be huge and the prospect of defining disease-related changes in the short term would be small. Fortunately, not all genetic variants operate independently. Alleles that lie close to each other on the chromosome are often found together more often than one would predict if they were segregating independently. This results from a variety of mechanisms but is most commonly the result of the historical development of mutations on a haplotype. New variants always occur in the context of already existing variants. These allelic associations will break down over time but that will depend on recombination between adjacent polymorphisms and, in some cases, selection. Because of the time frame necessary to separate these associations, many ancestral haplotypes exist widely throughout the genome. The relationship between polymorphic markers on the chromosome is referred to as linkage disequilibrium (LD) [9].

Those interested in the genetics of autoimmune disease will be familiar with the concept of LD as it exists within the HLA complex. The earliest studies of the relationship between autoimmune disease and HLA alleles suggested associations with HLA class I loci. As typing improved, it became evident that many of these associations related instead to class II alleles that reside several hundreds of kilobases away from the class I region. Fortunately for those using the HLA as a locus for disease genetics, the level of polymorphism in this region is extraordinarily high and there is extensive LD across the region. So, for example, the original association between HLA B8 and type I diabetes has now been shown to reflect the causative mutations within the HLA DQ region [10].

Some of the ancestral haplotypes extend over the entire region and show very high levels of conservation (i.e. A1, B8, DR3). This LD allowed the region to be associated with a large number of autoimmune diseases and, in some cases, even diseases that are nonimmune in nature but where genetic variation occurred on top of the existing ancestral haplotype. The association of the mutation in the HFE protein responsible for haemochromatosis with the HLA A3 haplotype is a clear example of this pattern of LD [11]. The rich disease gene mining that occurred by studying the HLA succeeded only because of the strong LD in the region, because none of the polymorphisms serologically studied initially were the variants responsible for disease.

Little is known about the mechanisms responsible for the persistent and strong LD in the HLA region. We do know that the alleles in the region are the subject of strong balancing selection. Clear evidence now exists for the role of these alleles in determining the response to infectious pathogens that have exerted powerful selective forces on human populations. For example, the severity of malaria infection seems to be determined by the presence of alleles of the HLA loci [12]. Recombination in this HLA region, detected in family studies, suggests that crossovers occur at a rate (2 cM/Mb) that is somewhat higher than the average around the genome.

Recent studies have attempted to study in detail the pattern of LD within a small portion of the MHC region [13]. In the region between DNA and TAP2, a region of approximately 200 kb, there are three regions of very strong LD that are broken up into discreet regions where LD is broken down. Sperm typing has been used to define the precise regions of meiotic recombination, and this revealed that such crossovers are highly clustered into five recombinational hot spots [14]. One of these hot spots (DNA 3) has a recombination frequency of 140 cM/Mb. In total, 94% of the crossovers recorded occurred within hot spots, although no sequence motifs defining such regions were recognized. Detailed analysis of this region by several groups has therefore defined regions of LD extending across distances of up to 100 kb that are interrupted by short recombinational hot spots [15,16].

Although much is known about patterns of LD within the MHC, the important outstanding question for human genetics is whether the same pattern of LD exists elsewhere in the genome. The large, dense set of SNPs now available is the resource necessary to undertake such studies. Predictions of LD across the genome have been unhelpful, largely because it is widely recognized that the patterns are heterogeneous. Studies of individual regions have suggested that, typically, LD often extends 60 kb from common alleles when tested with islands of sequences leading out from such common SNPs. In characterizing 19 such regions [14], the pattern of LD was greater than expected and there was evidence of significant variation from region to region. LD declined as distance grew from the central allele. Interestingly, this study established a correlation between the recombination rate within a genomic segment and the extent of LD in that region.

Although such studies provide some indication of what may be evident on a genome wide basis, more robust data comes from systematic approaches to haplotype mapping using large sets of markers across whole chromosomes. Such marker sets are widely available as a result of the SNP map. The first of these to be systematically tested covers chromosome 22 with a set of 1504 SNPs studied in the CEPH reference families (Dawson E *et al*., manuscript submitted). Again, significant heterogeneity in levels of LD were seen, with large regions >700 kb showing extensive LD while other regions revealed little evidence of LD, even between adjacent high-frequency alleles. Systematic study of LD on chromosome 21 has similarly been performed using radiation hybrids and oligonucleotide arrays, and it has revealed similar areas of limited haplotype diversity [17].

It would appear, therefore, that the situation seen in the MHC region with regard to LD is not unique and that extensive regions of LD exist throughout the genome, but that significant heterogeneity also exists between regions. A central, remaining question is how such regions occur and are maintained. Population history remains important and large areas of LD may reflect the lack of opportunity for recombination because the time frame has been short since the founder haplotype occurred or since an evolutionary bottleneck occurred. As in the MHC, the role of recombination rates and sites of recombinational activity may also be important. Selection also cannot be discounted in this process. Particularly in regions where gene products are heavily selected and potentially interactive, this may determine the coexistence of neighbouring alleles. Little is known about the sequences associated with high recombination rates, nor is there any good data on the sequences most associated with high LD.

Importantly, these data suggest that haplotypes are widespread throughout the genome and, whatever their mechanisms, may be valuable tools for moving into association studies in common disease. By carefully choosing markers from regions where LD is strong, it should be possible to undertake anonymous association strategies with the minimal number of SNPs. This approach should make non-hypothesis-based association strategies viable in the short term, allowing genetic effects such as those seen in the HLA region to be readily recognized and validated.

The use of such SNP tools holds great promise, but is not without its own complications. Disease-related genomic regions will be easier to detect in regions of strong and extensive LD. Once detected, however, it becomes challenging to detect the exact locus and polymorphism that accounts for disease susceptibility. This obstacle remains an issue for most HLA disease associations because the strong LD makes difficult the characterization of the role of any single polymorphism in the region where multiple variants contribute identical genetic information. This problem has arisen repeatedly in the HLA and has also been seen in other regions, such as the region around the insulin locus on chromosome 11 in the study of type I diabetes [18] and, more recently, in the study of the role of the cytokine cluster on 5q21 in Crohn's disease [19]. So although the presence of strong LD facilitates the identification of disease-associated regions, it makes the characterization of disease mutations or polymorphisms more difficult.

Another important issue to consider is that patterns of LD vary significantly between different population groups. This is potentially a mechanism whereby LD can be broken down to allow disease polymorphisms to be mapped. Such an approach was used to describe the role of individual loci in type I diabetes susceptibility utilizing the differential HLA haplotypes seen in Africa and Western Europe [20]. Ethnic heterogeneity in LD, however, also proves challenging in interpreting disease association data. Any bias in ethnic sampling between patients and controls can give rise to inappropriate assumptions about responsible loci and levels of risk that they confer.

The pattern of LD that is determined by population history can also be seriously influenced by the use of founder populations. The advantages of such approaches have been emphasized for studying complex traits. Populations with strong founder effects have obvious advantages in that they share significant genetic determinants and hence the locus and allelic heterogeneity that complicate the study of these diseases are reduced.

Population history determined the benefits derived from LD. When Eaves *et al*. characterized Finnish and Sardinian isolates, and compared their LD pattern in a region of chromosome 18 with Western European mixed populations, the levels of LD were not significantly different [21]. They suggest that this may reflect the entry of individual polymorphisms into the population through a number of different founders. The history of genetic isolates is therefore crucial for understanding the LD achieved.

Rare polymorphisms are likely to have arisen recently or from a single founder. If the variant is recent then the LD is probably strong and will represent the relatively short recombinational history of the haplotype. When the polymorphism is found at a higher frequency, it may have arisen from multiple founders and one would not expect the LD to be dramatically increased. The LD around a common variant will be increased only if it has arisen on a founder haplotype from a limited number of individuals, and hence this is likely to arise only in a small, isolated population. Overall, therefore, the role of founder populations in facilitating the search for common disease genes and polymorphisms is complex and the utility relates as much to the population history and age of the variants as it does to the presence of the concept of founders in the population.

Association strategies using SNPs in common disease

One of the most important applications that will arise from the availability of a large number of SNPs and LD on haplotype maps around the genome is the ability to systematically undertake association studies directed at identifying genetic determinants for common disease [22]. Association studies have been enormously successful at identifying some of the major loci involved in common disease. Associations between HLA alleles, first determined by serology and more recently by DNA typing, have identified a role for HLA gene products in determining susceptibility to a range of autoimmune diseases, including type I diabetes, rheumatoid arthritis, coeliac disease, multiple sclerosis and ulcerative colitis.

The success of these studies has, however, provided a false sense of security around the use of association strategies in mapping disease genes. The history of association studies in autoimmune disease since the identification of HLA associations has been limited. Many associations have been identified and then proven not to be reproducible in subsequent studies. Association strategies have therefore developed a bad reputation for producing false-positive results. Nevertheless, it is clear that, when properly applied, association strategies have potentially much greater power to detect genetic contributions to common complex diseases than that available using linkage studies. Not only is the statistical power to detect the genetic contribution to disease greater in association studies [23], it is also an essential component of the proof necessary to implicate a DNA variant within a linkage region as mediating disease susceptibility or pathogenesis.

An analysis of the potential pitfalls linked with association strategies is helpful, particularly at this time when they are likely to be more widely applied with the increasingly available SNP resources. First and foremost among the problems surrounding association strategy is the failure to have a sufficiently large sample size to produce convincing statistical support for a hypothesis. Few, if any, genetic loci yet to be identified in autoimmune disease will have the strength of HLA associations. Many of the genetic determinants of autoimmune disease will be contributing relative risks of between 2 and 4; as a result, large numbers of patients will need to be studied to identify these effects. These effects will also be diluted by the considerable heterogeneity that may underlie these diseases and by the fact that most genetic associations originally described are unlikely to be primary associations, but simply represent LD with the functional DNA variant responsible for disease. Both locus and allelic heterogeneity may add complexity to this formula. Together these confounding factors mean that large sample sizes (many hundreds of cases and controls) need to be studied to detect significant effects.

Population stratification is often identified as a cause for false-positive association studies. Although this has often been suggested, it has seldom been documented with any degree of rigour. Nevertheless, it is important to identify control populations that most properly represent samples for comparison with the disease group. Several strategies can be applied for identifying control populations [22]. Age-matched and sex-matched controls from similar ethnic backgrounds will provide some degree of confidence in variation in allele frequencies between patients and controls. Alternatively, sampling widely from diverse ethnic populations may provide information on the normal range of allele frequencies in diverse population groups.

Family-based controls provide an opportunity to reduce population stratification [24]. Should the disease population vary substantially from these, it provides confidence that the allele may be genuinely contributing to the disease state. Multiple testing and repeated subgroup analysis are conventional errors in methodology often seen in association strategies in human disease. Such strategies have proved to be misleading in epidemiological studies of all kinds and, increasingly, have proved to be the cause of false-positive data in genetic association studies. Any analysis of subsets of the disease population gives rise to serious statistical problems and can only be corrected by large and robust replicated studies, where the hypothesis is defined before the study is undertaken.

The availability of a large number of SNPs has also given rise to a problem associated with multiple testing of different allelic variants in case and control populations. At its worst, this problem is seen in 'whole genome association' strategies whereby very large numbers of SNPs are typed in disease and control populations. For example, if 1000 random SNPs are utilized for such association studies, one would expect 50 associations to be found at a significance level of $P < 0.05$. Given the issues of power already discussed, this strategy on its own is clearly fraught with difficulties. The significance of an association is obviously enhanced by the characterization biologically of an important candidate in disease pathogenesis or by the availability of previous data that points to a role for a specific polymorphism or gene in disease. Again, however, *post hoc* identification of candidates (e.g. in the region of linkage) can be a misleading approach, which carries with it risks of data overinterpretation.

The availability of haplotype maps around the genome provides perhaps the best opportunity for large-scale association strategies in autoimmune disease and other common disorders [25–27]. Based on currently available data, it would appear that some significant component of the genome can be analysed by the relatively small number of markers covering the areas where LD is strong and the number of haplotypes is limited. Within carefully controlled studies it may therefore be possible to eliminate, by association strategies, large segments of the genome from contributing any significant degree of risk to complex diseases, while other regions, where strong disequilibrium exists, may be identified for future analysis.

While this strategy has the potential of ruling in or ruling out very large segments of the genome for genetic analysis, the regions that have been identified through LD mapping as having a role in common disease provide significant challenges for the characterization of individual disease variants. Regions where LD is strong are extremely difficult to break down to the level of single DNA variants. This is particularly evident in regions such as the HLA, where the search for individual DNA variants responsible for disease associations continues to challenge

immunogeneticists. Nevertheless, this appears to be a strategy that will yield important association data, at least in the short term.

Concluding remarks

The availability of a large number of SNPs widely dispersed throughout the genome is likely to greatly accelerate disease gene hunting in autoimmunity. SNPs are thus not always independent, and the use of LD may facilitate the systematic search for associations. Before this happens, however, it is important to be clear about the methodological issues that have limited the effectiveness of previous simple association strategies. Only then will definitive genetic association data emerge.

References

1. Nakamura Y, Koyama K, Matsushima M: **VNTR (variable number of tandem repeat) sequences as transcriptional, translational or functional regulators.** *J Hum Genet (Jpn)* 1998, **43**:149-152. [general reference]
2. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M: **Genomewide scans of complex human diseases: true linkage is hard to find.** *Am J Hum Genet* 2001, **69**:936-950. [key review]
3. Mullikin JC, Hunt SE, Cole CG, Mortimore BJ, Rice CM, Burton J, Matthews LH, Pavitt R, Plumb RW, Sims SK, Ainscough RM, Attwood J, Bailey JM, Barlow K, Bruskiewich RM, Butcher PN, Carter NP, Chen Y, Clee CM, Coggill PC, Davies J, Davies RM, Dawson E, Francis MD, Joy AA, Lamble RG, Langford CF, Macarthy J, Mall V, Moreland A, Overton-Larty EK, Ross MT, Smith LC, Steward CA, Sulston JE, Tinsley EJ, Turney KJ, Willey DL, Wilson GD, McMurray AA, Dunham I, Rogers J, Bentley DR: **An SNP map of human chromsome 22.** *Nature* 2000, **407**:516-520. [general reference]
4. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D: **A map of human genome sequence variation containing 1.42 single nucleotide polymorphisms.** *Nature* 2001, **409**:928-933. [general reference]
5. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES: **An SNP map of the human genome generated by reduced representation shotgun sequencing.** *Nature* 2000, **407**:513-516. [general reference]
6. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280**:1077-1082. [general reference]
7. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22**:139-144. [key review]
8. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273**:1516-1517. [key review]
9. Hartl DL, Clark AG: *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates; 1997. [book]
10. She JX: **Susceptibility to type I diabetes: HLA-DQ and DR revisited.** *Immunol Today* 1996, **17**:323-329. [key review]
11. Cullen LM, Anderson GJ, Ramm GA, Jawinska EC, Powell LW: **Genetics of hemochromatosis.** *Annu Rev Med* 1999, **50**:87-98.[key review]
12. Hill AV: **Genetic susceptibility to malaria and other infectious diseases: from the MHC to the whole genome.** *Parasitology* 1996, **112**:S75-S84. [general reference]
13. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex.** *Nat Genet* 2002, **29**:217-222. [general reference]
14. 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. [general reference]
15. Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO: **Extent and distribution of linkage disequilibrium in three genomic regions.** *Am J Hum Genet* 2001, **68**:191-197. [general reference]
16. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF: **Haplotype variation and linkage disequilibrium in 313 human genes.** *Science* 2001, **293**:489-493. [general reference]
17. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR: **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21.** *Science* 2001, **294**: 1719-1723. [general reference]
18. Bennett ST, Lucassen AM, Gough SC, Powell EE, Undlien DE, Pritchard LE, Merriman ME, Kawaguchi Y, Dronsfield MJ, Pociot F, Nerup J, Bouzekri N, Cambon-Thomsen A, Rønningen KS, Barnett AH, Bain SC, Todd JA: **Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus.** *Nat Genet* 1995, **10**:378-380. [general reference]
19. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, Kulbokas EJ, O'Leary S, Winchester E, Dewar K, Green T, Stone V, Chow C, Cohen A, Langelier D, Lapointe G, Gaudet D, Faith J, Branco N, Bull SB, McLeod RS, Griffiths AM, Bitton A, Greenberg GR, Lander ES, Siminovitch KA, Hudson TJ: **Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease.** *Nat Genet* 2001, **29**:223-228. [general reference]
20. Lampis R, Morelli L, Congia M, Macis MD, Mulargia A, Loddo M, De Virgiliis S, Marrosu MG, Todd JA, Cucca F: **The inter-regional distribution of HLA class II haplotypes indicates the suitability of the Sardinian population for case–control association studies in complex diseases.** *Hum Mol Genet* 2000, **9**:2959-2965. [general reference]
21. Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA: **The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes.** *Nat Genet* 2000, **25**:320-323. [general reference]
22. Cardon LR, Bell JI: **Association study designs for complex diseases.** *Nat Rev Genet* 2001, **2**:91-99.[key review]
23. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405**:847-856.[key review]
24. Spielman RS, McGinnis RE, Ewens WJ: **Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus.** *Am J Hum Genet* 1993, **52**:506-516. [general reference]
25. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10**:1435-1444. [general reference]
26. Xiong M, Guo SW: **Fine-scale genetic mapping based on linkage disequilibrium: theory and applications.** *Am J Hum Genet* 1997, **60**:1513-1531. [general reference]
27. Freimer NB, Reus VI, Escamilla MA, McInnes LA, Spesny M, Leon P, Service SK, Smith LB, Silva S, Rojas E, Gallegos A, Meza L, Fournier E, Baharloo S, Blankenship K, Tyler DJ, Batki S, Vinogradov S, Weissenbach J, Barondes SH, Sandkuijl LA: **Genetic mapping using haplotype, association and linkage methods suggests a locus for severe bipolar disorder (BPI) at 18q22-q23.** *Nat Genet* 1996, **12**:436-441. [general reference]