doi:10.1093/brain/awn081 *Brain* (2008), **131**, 3118–3131

REVIEW ARTICLE

The complex genetics of multiple sclerosis: pitfalls and prospects

Stephen Sawcer

University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's, Hospital, Hills Road, Cambridge, CB2 2QQ, UK

Correspondence to: Stephen Sawcer, University of Cambridge, Department of Clinical Neurosciences, Addenbrooke's, Hospital, Hills Road, Cambridge, CB2 2QQ, UK
E-mail: sjs10l6@mole.bio.cam.ac.uk

The genetics of complex disease is entering a new and exciting era. The exponentially growing knowledge and technological capabilities emerging from the human genome project have finally reached the point where relevant genes can be readily and affordably identified. As a result, the last 12 months has seen a virtual explosion in new knowledge with reports of unequivocal association to relevant genes appearing almost weekly. The impact of these new discoveries in Neuroscience is incalculable at this stage but potentially revolutionary. In this review, an attempt is made to illuminate some of the mysteries surrounding complex genetics. Although focused almost exclusively on multiple sclerosis all the points made are essentially generic and apply equally well, with relatively minor addendums, to any other complex trait, neurological or otherwise.

Keywords: multiple sclerosis; genetics; association; linkage

Abbreviations: GWAS = Genome-Wide Association Study; HLA = human leucocyte antigens; IL2R = interleukin-2 receptor; IL7R = interleukin-7 receptor; IMAGEN = International MHC and Autoimmunity Genetics Network; IMSGC = International Multiple Sclerosis Genetics Consortium; LD = linkage disequilibrium; MHC = major histocompatibility complex; MAF = minor allele frequency; NIMH = National Institutes of Mental Health; nsSNPs = non-synonymous single nucleotide polymorphism; RAF = risk allele frequency; SNP = single nucleotide polymorphism; WTCCC = Wellcome Trust Case Control Consortium

Received March 10, 2008. Revised March 27, 2008. Accepted April 2, 2008. Advance Access publication May 18, 2008

Evidence for the influence of genetics

Most, if not all, common diseases are characterized by an increased frequency in the relatives of affected individuals, and multiple sclerosis is no exception. Amongst white individuals living in temperate regions, the prevalence of multiple sclerosis is typically 1/1000, yet 15–20% of patients report a family history of the disease, a rate which is significantly more than would be expected by chance (Compston and Coles, 2002). Several carefully conducted, population-based studies of familial recurrence risk have been performed which have confirmed and quantified the increased risk of the disease in the relatives of affected individuals (Sadovnick *et al.*, 1988; Robertson *et al.*, 1996; Carton *et al.*, 1997). This familial clustering can be usefully summarized in terms of λ_s; the relative risk of the disease seen in the siblings of affected individuals as compared to that seen in the general population (Risch, 1990). In multiple sclerosis this risk ratio takes a value of ~15 (Sawcer, 2006), and of course reflects the combined effects of all shared aetiological influences, both genetic and environmental (Guo, 2002). Teasing apart the relative contribution of these alternate influences requires the analysis of informative groups and many such studies have been performed in multiple sclerosis. A significant excess of concordance in monozygotic (identical) as compared to dizygotic (non-identical) twins is a virtually universal finding (Mumford *et al.*, 1994; Willer *et al.*, 2003; Hansen *et al.*, 2005; Ristori *et al.*, 2006; Islam *et al.*, 2006), with the only exceptions being studies which lacked the power to show any difference as significant (French Research Group on Multiple Sclerosis, 1992). In a study of individuals with multiple sclerosis who had been adopted in early life, researchers found no excess risk of the disease in the adopting family but the expected excess risk within

the natural (genetically related) family (Ebers *et al.*, 1995). Couples where both the husband and the wife are affected with the disease (so-called conjugal pairs) do not occur more frequently than would be expected by chance, although the risk of the disease in their offspring is greater than if just one parent is affected (Robertson *et al.*, 1997; Ebers *et al.*, 2000). The risk of the disease in half-siblings is approximately half the risk seen in full siblings, regardless of whether they are raised together or apart (Ebers *et al.*, 2004), while there is no excess risk in stepsiblings (Dyment *et al.*, 2006), who are of course genetically unrelated. In summary, these data suggest that living with someone who has, or who will eventually develop, multiple sclerosis has little or no effect on your risk of developing the disease unless you are genetically related to them, in which case your risk increases with the degree of relatedness (Dyment *et al.*, 2006). This is not to imply that environmental factors have no role, only that they seem to exert their effects mainly at a population level with the micro-environmental differences between families within a given population seeming to be of relatively little importance (Dyment *et al.*, 2006). Although these data confirm that genetic factors are unequivocally relevant in multiple sclerosis, large extended families containing multiple affected individuals in multiple generations are extremely uncommon (Willer *et al.*, 2007). Most families contain no more than two or three affected individuals and no clear mode of inheritance can be inferred from segregation analysis (Compston *et al.*, 2006).

The epidemiology of multiple sclerosis continues to be scrutinized and interesting nuances will no doubt continue to emerge. However, it is important to recognize the limitations of this approach. These studies are extremely difficult to perform and frequently subject to confounding and bias. Moreover, although multiple sclerosis is a 'common' disease in neurological practice, the relatively modest frequency of the disease in the population as a whole (1/1000) means that even huge population-based studies can only provide crude estimates for recurrence risks, and other basic epidemiological parameters, such as age-specific incident rates and life time risk, all of which come with large confidence intervals. It seems likely that most, if not all, of any apparent inconsistencies in the epidemiology of multiple sclerosis stem from the variability inherent in underpowered studies. In light of these considerations, it is easy to see why predictable effects, such as the Carter effect, have only inconsistent support (Hupperts *et al.*, 2001; Ebers *et al.*, 2004; Kantarci *et al.*, 2006; Herrera *et al.*, 2007) (see Supplementary material section 1). In a complex disease like multiple sclerosis where epidemiological parameters are impossible to measure reliably and multiple potentially conflicting effects are likely to exist, it seems unlikely that epidemiological analysis will ever provide any major insights. In short, epidemiological analysis has convincingly shown us that genetic factors are relevant but lacks any power to illuminate the nature or extent of these factors beyond indicating that multiple genes are involved (Wang *et al.*, 2005; Lindsey, 2005).

Early success

Initial attempts to identify genes influencing susceptibility to multiple sclerosis were highly successful and quickly identified the now well-established relevance of the Major Histocompatibility Complex (MHC). Unfortunately, this early success has been followed by several decades of frustration in which no other undeniably relevant loci emerged until 2007. Before contemplating why it has been so difficult to map non-MHC loci, it is worth revisiting the MHC association, its discovery and subsequent dissection.

Association between the Human Leukocyte Antigens (HLA) and multiple sclerosis was first identified in 1972. Using cell culture-based methods researchers from California found association with HLA-A3 (Naito *et al.*, 1972) while others from Denmark found association with HLA-B7 (Jersild *et al.*, 1972). The following year the same Danish group also established association with DR2 (Jersild *et al.*, 1973). The nomenclature used to describe HLA is complex and has evolved considerably over the years. At the time of these original discoveries very different designations were used, such that the phenotypes associated with the HLA-A3, HLA-B7 and DR2 alleles were known respectively as HL-A3, HL-A7 and LD-7a. It was quickly realised that these were not independent associations but were rather a reflection of linkage disequilibrium (LD, see Supplementary material section 2) between the corresponding alleles and association of the disease with a haplotype including these alleles (Compston *et al.*, 1976; Terasaki *et al.*, 1976). The molecular genetic dissection of these associations began in 1984 when Cohen *et al.* (1984) used analysis of restriction fragment length polymorphisms to directly establish association with the HLA-DR2 allele (Cohen *et al.*, 1984). Over the years, technology has improved and the resolution of the associated alleles has been refined (Vartdal *et al.*, 1989; Olerup and Hillert, 1991).

All of these associated HLA genes lie in the MHC, a gene-dense region of the genome characterized by extensive LD and extreme levels of polymorphism (Horton *et al.*, 2004). In light of these features, it is not unsurprising to find that many variants from other genes in this region also show association with multiple sclerosis (Lincoln *et al.*, 2005; Yeo *et al.*, 2007). The modest levels of LD between the class I region (containing the HLA-A and HLA-B genes) and the class II region (containing the DRB1 and DQB1 genes) enabled researchers to quickly establish that association primarily derived from the class II region (Compston *et al.*, 1976; Terasaki *et al.*, 1976). However, the more extensive LD between DRB1 and DQB1 made it much more difficult to refine which of these genes was primarily responsible for the association. Studying African American patients, who have less intense LD between DRB1*1501 and DQB1*0602, Oksenberg *et al.* (2004) provided the first

convincing evidence that the primary association was with the DRB1 gene, an observation which has been confirmed in subsequent studies in large cohorts of patients of European descent (Yeo *et al.*, 2007). Because of further evolution in the nomenclature of HLA genes what was previously called DR2, is these days referred to as DR15; the DRB1*1501 allele is the most common sub type of DR15 seen in white Europeans.

Looking back at some of these original studies in light of current knowledge is highly informative. Even though the extent of linkage disequilibrium between HLA-A3 and HLA-DRB1*1501 is modest ($D' = 0.3$, $r^2 = 0.14$) the original study by Naito *et al.* (1972), which included 94 cases and 871 controls, had >50% power to identify association with A3 at the 5% level. This early study thus illustrates well the principle that genuine associations can indeed be identified by typing markers in LD with real effects even when the level of LD is modest. Of course the saving grace for Naito *et al.* (1972) was the strength of the association with the DRB1*1501 allele and their use of a large cohort of controls. In the study by Compston *et al.* (1976), a class II locus was considered directly and nominally significant association was confirmed using just 83 cases and 32 controls.

It is now well established that the association of multiple sclerosis with the DRB1*1501 allele is almost ubiquitous, the relevance of this allele having been confirmed in virtually every population tested (Compston *et al.*, 2006). The fact that other MHC haplotypes also influence susceptibility is well established (Marrosu *et al.*, 1998) and recent data indicate that the risk associated with *1501 may be modified depending upon which other MHC haplotype is carried in the heterozygous state (Dyment *et al.*, 2005; Barcellos *et al.*, 2006; Ramagopalan *et al.*, 2007). However, it is unclear whether these additional signals stem primarily from the DRB1 gene or from the effects of other MHC loci. Many researchers have found evidence supporting the existence of an independent signal from the class I region (Fogdell-Hahn *et al.*, 2000; Marrosu *et al.*, 2001; de Jong *et al.*, 2002; Rubio *et al.*, 2002; Harbo *et al.*, 2004; Yeo *et al.*, 2007) although not all (Lincoln *et al.*, 2005). Again, this apparent inconsistency is not unexpected. Establishing the presence of additional susceptibility loci located close to a primary locus is complex especially in the presence of prominent LD and likely allelic heterogeneity (Koeleman *et al.*, 2000). Once correction for the effects of LD with the DRB1*1501 allele are made, the residual power in even the largest of these studies is modest (Dyment *et al.*, 2005; Lincoln *et al.*, 2005; Barcellos *et al.*, 2006; Yeo *et al.*, 2007). A role for the DRB1∗03 haplotype seems beyond doubt (Dyment *et al.*, 2005; Barcellos *et al.*, 2006; Yeo *et al.*, 2007) and Sardinian data would suggest that this association most likely stems from the DRB1 gene (Marrosu *et al.*, 2001). Beyond this it is clear that the MHC contains further signals but their nature and origins are as yet unresolved.

By cataloguing variation in the MHC through the re-sequencing of specific haplotypes (Allcock *et al.*, 2002; Horton *et al.*, 2008), and empirically establishing the complex patterns of LD across the region (Miretti *et al.*, 2005), it has been possible to establish a comprehensive panel of haplotype tagging single nucleotide polymorphisms (SNPs) (de Bakker *et al.*, 2006). These SNPs are currently being typed in multiple sclerosis and a number of other autoimmune diseases as part of the International MHC and Autoimmunity Genetics Network project. Hopefully, these systematic fine-mapping efforts will help to unravel this complex association, although it can be anticipated that large sample sizes will be needed to confirm the findings emerging from this project.

The rest of the genome

Outside the MHC, the genetic analysis of multiple sclerosis has been considerably less successful, with no consistent findings emerging until very recently. This lack of any convincing progress has been a source of great frustration, and the inconsistency in early claims has rightly been criticized (Hirschhorn *et al.*, 2002). It is now clear that two main issues have confounded the identification of relevant genes—the modest size of effects attributable to individual loci (Ioannidis *et al.*, 2006) and a failure to correctly allow for the statistical consequences that result from the enormous size of the genome (Ioannidis, 2003). The search for the genes of relevance in multiple sclerosis has reasonably been likened to searching for a handful of rather small needles in a very large haystack (Hensiek *et al.*, 2003*b*).

The needles

Obviously we cannot know *a priori* what effects on risk will be conferred by individual susceptibility alleles, nor can we know their frequency or mode of inheritance. However, linkage analysis has provided us with invaluable guidance regarding an upper limit on these effects sizes, which researchers cannot afford to ignore.

The fact that association with the MHC can reliably be detected with modest resources (c 100 cases and 100 controls) and yet only accounts for a fraction of the heritability seen in multiple sclerosis meant that in the late 1980s and early 1990s there was an expectation that non-MHC loci would be rather easy to find. At this time, there was a feeling that perhaps susceptibility to multiple sclerosis might be determined by just a handful of effects similar, or perhaps even larger, to that conferred by the MHC. Coincident with this the human genome project reached the point where systematic whole-genome screening for linkage became possible (see section 3 of Supplementary material). In 1996, the results of whole-genome screens for linkage to multiple sclerosis from the UK, the US and Canada were published back to back (Ebers *et al.*, 1996; Haines *et al.*, 1996; Sawcer *et al.*, 1996). Each of these

studies was based on ~100 affected sib pairs and employed 300–400 microsatellite markers. Subsequently similar studies were performed in multiplex families from Finland (Kuokkanen *et al.*, 1997), Sardinia (Coraddu *et al.*, 2001), Italy (Broadley *et al.*, 2001), Scandinavia (Akesson *et al.*, 2002), Australia (Ban *et al.*, 2002) and Turkey (Eraksoy *et al.*, 2003), and in addition each of the original three groups extended their analysis using further families and more microsatellite markers (Hensiek *et al.*, 2003*a*; Dyment *et al.*, 2004; Kenealy *et al.*, 2004). Interestingly, none of these studies identified any statistically significant linkage, not even in the region of the MHC. Attempts at meta-analysis were no more successful, although linkage to the MHC just reached genome-wide significance in some of these studies (Ligers *et al.*, 2001; GAMES and the Transatlantic Multiple Sclerosis Genetics Cooperative, 2003). Contemplating the reasons for this disappointing lack of linkage, the International Multiple Sclerosis Genetics Consortium (IMSGC) identified a number of issues which might have confounded these studies (IMSGC, 2004) and in an attempt to correct for these re-screened the genome for linkage using a dense map of SNPs in families from Australia, Scandinavia, the US and the UK (IMSGC, 2005). In the final analysis, this substantially larger study included data from 4506 SNPs typed in 730 multiplex families which between them provided almost 1000 affected relative pairs. The increased power provided by this screen in comparison with its predecessors is evident from the overwhelming evidence for linkage found in the MHC region where a lod score of 11.7 was observed (IMSGC, 2005). Once again, however, no other region of statistically significant linkage was apparent. The comprehensive marker map used in this study makes it virtually impossible that any signals of a magnitude similar to that attributable to the MHC could have been missed. As with the previous studies the number of suggestive linkage peaks was significantly greater than would have been expected by chance alone (IMSGC, 2005), indicating that there is excess allele sharing but providing no clear guide as to the location of relevant genes.

Although these linkage data provide no useful information concerning the location of non-MHC susceptibility loci the observed allele sharing does provide useful guidance concerning the size of effects attributable to such loci (Risch, 1990). Employing the approach suggested by Risch and Merikangas (1996), and remembering that the observed allele sharing is expected to provide a significantly inflated estimate of effect size (Goring *et al.*, 2001), it is straightforward to show that common non-MHC risk alleles are highly unlikely to increase risk by more than a factor of 2.0. Under these circumstances, it is clear that further linkage analysis is almost certain to be unrewarding since the number of sib pair families necessary to demonstrate significant linkage is impractically large (Risch and Merikangas, 1996) (see Supplementary material section 3). Fortunately, association-based studies are significantly more powerful and thus provide a means to identify genes exerting effects which fall below the resolution of linkage (Risch and Merikangas, 1996). However, even the most optimistic estimates of effect size consistent with the available linkage data indicate that association studies will need to involve at the very least 500 cases and 500 controls (Sawcer, 2006). Since most of the published literature regarding the genetics of multiple sclerosis has been based on significantly smaller numbers one corollary of this is that almost all previous studies have been seriously underpowered. There are virtually no loci, with the possible exception of APOE (Burwick *et al.*, 2006), where published studies have been adequately powered to confidently exclude the possibility of a meaningful effect. It seems highly likely that many of the entirely plausible candidates considered to have been excluded on the basis of the absence of any consistent evidence to date will eventually emerge as genuinely relevant in the disease. Coupling this limited effect size with the fact that for most genes no more than a tiny fraction of the variation has ever been tested, it is clear that few if any genes have received a thorough analysis. It is surely the virtual absence of any power that is responsible for nearly all the apparent inconsistency in the literature concerning the genetics of complex diseases such as multiple sclerosis (Lohmueller *et al.*, 2003).

Of course, it remains possible that a large extended family in which a rare more penetrant allele is segregating might be found, and that the identification of such an allele might be informative regarding the pathogenesis of the disease, much as the identification of mutations in the alpha-synuclein gene has been informative regarding the pathogenesis of Parkinson's disease (Polymeropoulos *et al.*, 1997). However, analysis of the few such families so far reported has failed to identify any significant linkage let alone any relevant loci (Dyment *et al.*, 2002; Modin *et al.*, 2003). These larger families are characterized by multiple affected siblings rather than multiple affected generations, rarely show the greater degree of consistency in phenotype that would be expected but instead show an increased frequency of DRB1*1501 carriage, the reverse of what would be expected if a non-MHC locus were primarily responsible for the disease (Willer *et al.*, 2007). It also remains possible that some otherwise rare alleles of higher penetrance might have drifted through a genetic bottleneck and thereby become frequent in a population isolate. However, given the surprising extent of identity by descent seen in apparently unrelated individuals (Frazer *et al.*, 2007), it is hard to imagine that there will be much power to separate such alleles through homozygosity mapping.

The haystack

Although there are no clear data regarding the genetic architecture underlying susceptibility to multiple sclerosis considerable progress has been made concerning the nature and extent of genetic variation in the human population in general (Frazer *et al.*, 2007). The human genome is roughly

3 billion base pairs long (Human Genome Project, 2004) and is on average 99.9% identical between any two individuals (International HapMap Project, 2003). Although the total number of variants in the human population runs into billions, the majority of these are vanishingly rare (Kruglyak and Nickerson, 2001) and as a result most (~90%) of the differences between any two individuals is attributable to common variants, where both the alleles are seen in at least 1% of the population (Wang *et al.*, 2005). In light of these observations, two competing theories regarding the nature of susceptibility to complex disease have emerged. The first, the so-called common disease/common variant hypothesis (Reich and Lander, 2001; Pritchard and Cox, 2002), holds that susceptibility to common disease is determined by a few common variants with low penetrance, while the second, the heterogeneity (or multiple rare variant) hypothesis (Smith and Lusis, 2002), holds that the notion of a common disease is essentially a misnomer and that in fact such 'diseases' are in reality a collection of genetically distinct conditions each determined by a rare variant of higher penetrance. Although these two hypotheses represent opposite extremes, they are not mutually exclusive and it has been argued that perhaps the most parsimonious expectation is that both will be involved to some extent (Wang *et al.*, 2005), with some genuine heterogeneity but the bulk of the disease being determined by common variants of low penetrance. It is possible to estimate the number of variants that might account for a disease like multiple sclerosis (Lindsey, 2005; Yang *et al.*, 2005); under the common disease/common variant hypothesis just 20–100 common alleles each with a modest Genotypic Relative Risk (GRR) (in the order of 1.2–1.5) would be sufficient, while under the multiple rare variant model many hundreds if not thousands of rare alleles would be required even if the GRR attributable to each was considerable (in the order of 10–20) (Yang *et al.*, 2005). [See Risch and Merikangas (1996) and the Supplementary material for the definition of GRR]. Following this logic we might expect that perhaps 100 common variants exerting modest effects on risk (i.e. of low penetrance) are involved in determining susceptibility to multiple sclerosis, and might also expect that a small portion of the disease will turn out to have a distinct genetic basis related to rarer rather more penetrant alleles. Given that there are estimated to be some 10 million common variants in the human population as a whole (Kruglyak and Nickerson, 2001) we can see that the odds that any randomly selected common variant is relevant in multiple sclerosis is ~100 000:1 against (assuming that there is no significant LD between the various risk alleles).

By calculating the power of a study to identify any particular level of significance (Purcell *et al.*, 2003) and using the above estimate for the prior odds we can determine the odds that a result with any particular level of significance is a true positive (see Supplementary material section 4). Figure 1 shows the posterior odds for studies of differing size assuming that the risk alleles relevant in multiple sclerosis are common (frequency 10%) and have a GRR of 1.3 (under a multiplicative model).

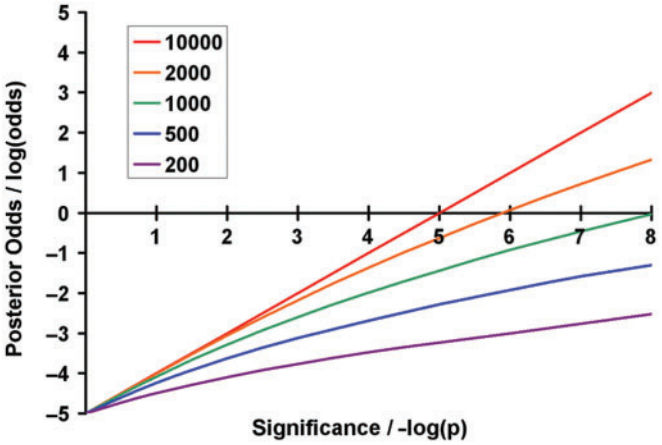

Fig. 1 Posterior Odds that a result is true, assuming risk alleles with a frequency of 10% and a Genotype Relative Risk (GRR) of 1.3 and a multiplicative model. This figure indicates the posterior odds that a result is true (plotted on a log scale on the *y*-axis) against the significance of the result (plotted as -log of the *P*-value on the *x*-axis). Five sample sizes are listed in the legend, in each the number of cases and controls are equal, the 200 line thus indicates the posterior odds for a study involving 200 cases and 200 controls and so on. Power was calculated using the on-line genetic power calculator (Purcell *et al.*, 2003).

From this figure, we can see that *P*-values in the range of 5–0.1% will virtually always be false positives, even in well-powered studies. This primarily occurs because the prior odds are so extreme that it remains more likely that this level of significance has arisen by chance in an unassociated marker than that we happen to have considered a genuinely associated marker. As the *P*-value becomes more extreme the probability of seeing such a result by chance alone is reduced (by definition) and therefore it becomes increasingly likely that the result is a true positive. While this is intuitively expected it is perhaps counter-intuitive to see that the smaller the study (i.e. the less power in a study) the greater the level of significance needs to be before a result becomes more likely to be true than false (WTCCC, 2007). For studies involving 1000 cases and 1000 controls we can see that only *P*-values $<10^{-8}$ are more likely to be true than false. In less well powered studies, such as those using just 200 cases and 200 controls, even such extreme *P*-values remain 100–1000 times more likely to be false than true. The Wellcome Trust Case Control Consortium (WTCCC) proposed this Bayesian approach and through similar reasoning have recommended that association studies in complex disease should involve at least 2000 cases and 2000 controls, in which case *P*-value of $<5 \times 10^{-7}$ will more often be true than false (WTCCC, 2007). Comparing the different curves in Fig. 1 it is clear that a considerable amount is gained by increasing the sample size from 200 to 2000, while, for effects of this size, relatively little is gained by

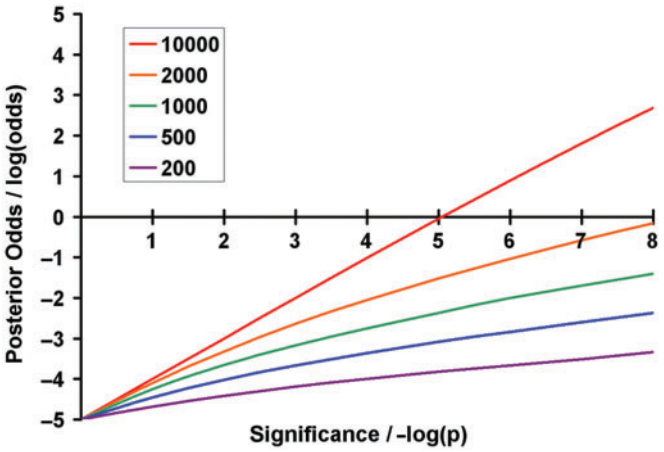

Fig. 2 Posterior Odds that a result is true, assuming risk alleles with a frequency of 10%, a GRR of 1.2 and a multiplicative model. The axes and samples sizes are as in Fig. 1. Power was calculated using the on-line genetic power calculator (Purcell *et al*., 2003).

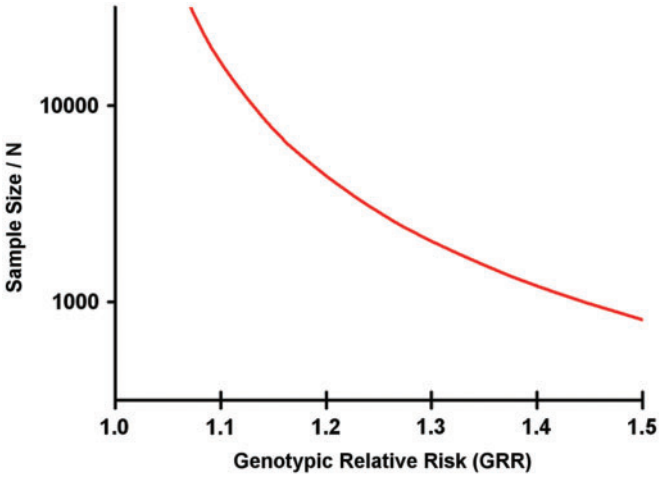

Fig. 3 Required sample size as a function of the GRR conferred by susceptibility allele. The sample size required is plotted on a log scale. The sample size indicates the number of cases required to ensure that results with a *P*-value of $<5 \times 10^{-7}$ are twice as likely to be true as false, assuming an equal number of controls, a multiplicative model and risk allele frequency of 10%. Sample sizes were calculated using the on-line genetic power calculator (Purcell *et al*., 2003).

increasing the sample size from 2000 to 10 000. However, for effects of smaller magnitude the value of large sample sizes becomes clear (Fig. 2).

These figures illustrate the need for adequate power. It is only when studies have sufficient power that we can rely on the prediction that *P*-values of $<5 \times 10^{-7}$ will more often be true than false. Figure 3 shows the sample size required in order to ensure that *P*-values of $<5 \times 10^{-7}$ will indeed be more often true than false in terms of the GRR conferred by susceptibility loci.

In the analysis presented above, it has been assumed that the variant chosen for study has been selected at random

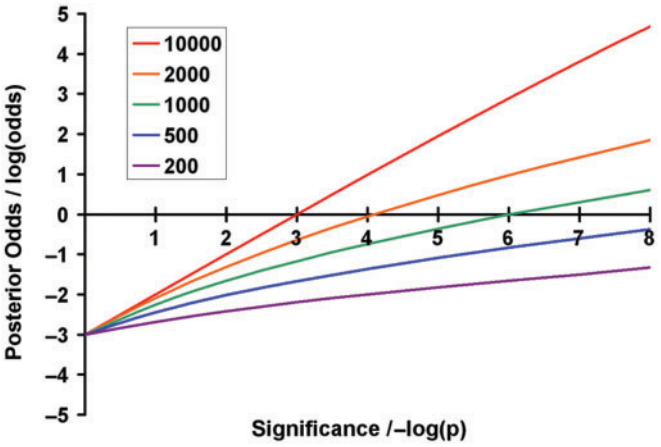

Fig. 4 Posterior odds that a result is true, assuming candidate risk alleles with a frequency of 10%, a GRR of 1.2, a multiplicative model and prior odds of 1000 : 1. The axes and samples sizes are as in Fig. 1. Power was calculated using the on-line genetic power calculator (Purcell *et al*., 2003).

from amongst the full list of common variation. In practice, researchers do not do this but instead tend to use some prior knowledge or existing information to guide the selection of 'candidates'. Of course, the validity of these prior data cannot be known and assumptions used to guide the selection of candidates may be invalid. In this way, we can imagine that random selection represents the worst-case scenario. Information used to guide the selection of candidates may come from many different sources such as data from animal models, expression studies, biological or pathological analysis. For example, the evidence that multiple sclerosis is an autoimmune disease is overwhelming and makes any gene with immunological function a logical candidate. However, since perhaps a fifth of genes have an immunological function using this information to guide the selection of candidates would only improve the prior odds by a factor of 5 taking them from 100 000:1 to 20 000:1. These odds are certainly reduced but come at a price since the likelihood of discovering non-immunological genes of relevance has been greatly reduced. Since non-synonymous coding variants and variants in regulatory regions or splice sites are more likely to have a functional effect than variants in silent non-coding regions it has also been suggested that concentrating analysis on these more functional relevant variants could also improve the prior odds (Tabor *et al*., 2002). Using multiple available sources of information to guide the selection of candidates in a process known as genomic convergence has been suggested (Hauser *et al*., 2003) but even this comprehensive approach seems unlikely to improve the prior odds much beyond 1000:1 (Wacholder *et al*., 2004). Figure 4 shows the posterior odds for the study of an optimally selected candidate variant, a situation which might be considered the best-case scenario.

From this figure we can see that even for well-selected candidates studied in cohorts involving as many as 1000

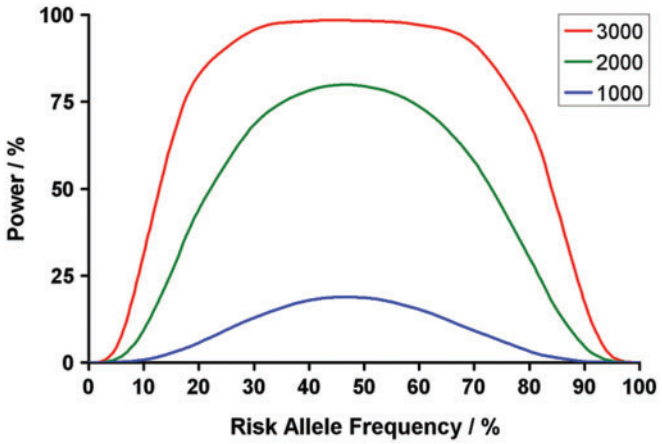

Fig. 5 Influence of risk allele frequency (RAF) on power to identify significant association (P-value $<5 \times 10^{-7}$). Power was calculated under the assumption that the susceptibility alleles have a GRR of 1.3 and a multiplicative model. Sample sizes are indicated in the legend. Power was calculated using the on-line genetic power calculator (Purcell *et al*., 2003).

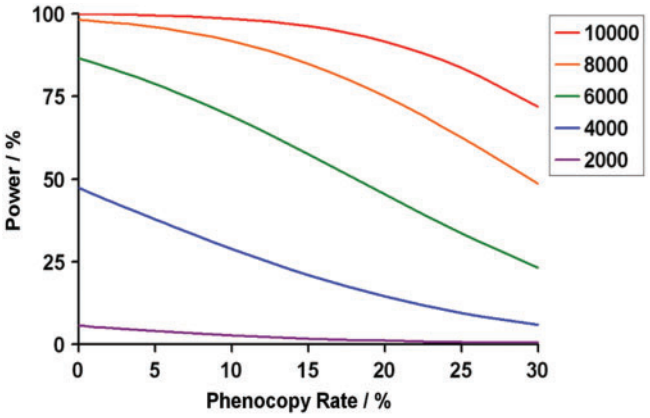

Fig. 6 Influence of phenocopy rate on power to identify significant association (P-value $< 5 \times 10^{-7}$). The phenocopy rate indicates the proportion of cases which have been misdiagnosed as having the disease when in fact they are controls. A phenocopy rate of 25% thus indicates that 1 in 4 of the cases is a misdiagnosis (or heterogeneity). Since the prevalence of multiple sclerosis is 1 per 1000 the impact of using unselected controls is minimal; for example in a cohort of 2000 unselected controls only two would be expected to be cases and thus the impact on power and posterior odds is imperceptible. Power was calculated under the assumption that the susceptibility alleles have a GRR of 1.3 and a multiplicative model. Sample sizes are indicated in the legend, note these are not all the same as in earlier figures. Power was calculated using the on-line power for association with error (PAWE) calculator (Gordon *et al*., 2002; Edwards *et al*., 2005).

cases and 1000 controls modest P-values (in the range of 5–0.1%) are still much more likely to be false positives than true. On the other hand, in a candidate gene study this number of samples is sufficient to ensure the reliability of more stringent P-values such as 5×10^{-7}.

It is expected that the frequency of risk alleles will vary from locus to locus so it is reasonable to enquire how this variable might influence the interpretation of results. Figure 5 shows the relationship between Risk Allele Frequency (RAF) and the power to identify meaningful association (P-value $<5 \times 10^{-7}$).

Inspection of Fig. 5 shows that the power drops precipitously as the minor allele frequency (MAF) falls below 20% (corresponding to RAF values of <20 or >80%) even in large study cohorts. Once the MAF falls below 5% there is virtually no power. On the other hand, for intermediate values of RAF there is relatively little variation in power.

The effects of heterogeneity on the ability to identify common susceptibility variants is also worthy of consideration. A degree of heterogeneity is to be expected (Wang *et al*., 2005) and the extent to which this and other sources of confounding, such as diagnostic inaccuracy, reduce the power to identify association is clearly relevant. Figure 6 indicates the consequences of including phenocopies in the case cohort (Gordon *et al*., 2002; Edwards *et al*., 2005).

From Fig. 6 it is clear that a surprisingly high level of phenocopy inclusion can be tolerated. This observation should not be interpreted as arguing for careless phenotyping, clearly power will be reduced every time a phenocopy is mistakenly included in a study and every effort should be made to keep this to a minimum. On the other hand, given that a degree of heterogeneity is expected it is important to realize that even if this amounted to as much as 10–15% of the disease it would still be possible to identify relevant common variants. Some evidence for heterogeneity in multiple sclerosis has been identified although this probably amounts to no more than 1–2% of the disease. In the early 1990s, it was realized that some patients with Leber's Hereditary Optic Neuropathy (an optic atrophy caused by specific mutations in mitochondrial DNA) developed a disease that, apart from the prominence of visual failure, was clinically and radiologically indistinguishable from multiple sclerosis (Harding *et al*., 1992; Riordan-Eva *et al*., 1995). This condition is, however, extremely rare. More recently investigators from the Mayo clinic (USA) have established that in some cases inflammatory demyelination of the central nervous system results from autoantibodies directed against the water channel aquaporin-4 (Lennon *et al*., 2004, 2005), these antibodies thereby providing the first biomarker distinguishing a pathogenically distinct subgroup of patients with CNS inflammatory demyelinating disease (Lennon *et al*., 2004). Although the main phenotypes associated with these antibodies, neuromyelitis optica (Devic's disease), recurrent myelitis and recurrent optic neuritis are highly distinctive at the clinical level (Wingerchuk *et al*., 2006) it is clear that the phenotype associated with these antibodies is expanding and we can expect that a proportion of cases satisfying clinical diagnostic criteria for multiple sclerosis (McDonald *et al*., 2001) will in fact turn out to have this distinct antibody-mediated disease. Outside these two rare conditions,

no significant heterogeneity has thus far been confirmed in multiple sclerosis.

Some investigators feel that primary progressive disease is a distinct condition and should be considered separately from relapse remitting disease, while others feel that this apparent distinction is just a reflection of the fact that the activity of the relapsing component of the disease is highly variable, being essentially absent in some cases and prominent in others. Detailed analysis of the natural history of the disease has shown that progression is essentially independent of relapse activity and indistinguishable between primary progressive and relapse onset cases (Compston, 2006; Confavreux and Vukusic, 2006*a*, *b*; Kremenchutzky *et al*., 2006). In the same way, pathological and radiological differences between primary progressive and relapsing onset disease are largely a reflection of relapse activity rather than being distinct. It seems likely that genetic factors will influence the course of multiple sclerosis, and there is evidence for a degree of concordance within multiplex families with respect to course (Hensiek *et al*., 2007). However, it also seems likely that in terms of susceptibility factors there will be rather more in common between primary progressive and relapsing disease than different, certainly there is no convincing evidence for any difference between these two groups in terms of the susceptibility factors thus far established.

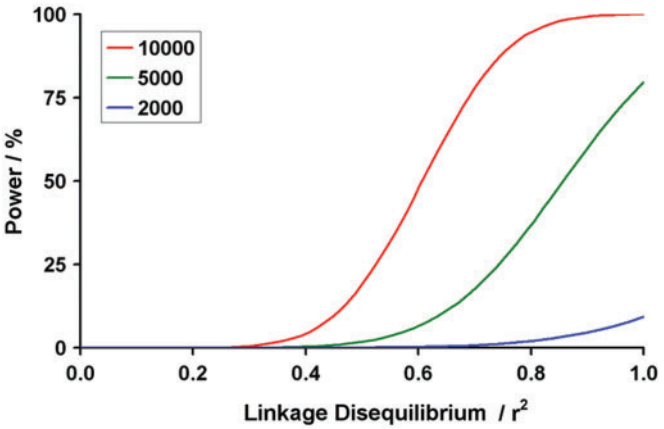

Fig. 7 Influence of linkage disequilibrium (LD) between the tested and causative variant on the power to identify significant association (P-value $<5 \times 10^{-7}$). There is a simple relationship between the extent of linkage disequilibrium and effective sample size (Wall and Pritchard, 2003), such that the product of r^2 (see Supplementary material section 2) and actual sample size indicates the sample size which would have yielded the same power if the causative variant had been directly typed. This simple relationship was used to calculate the effective sample size at each level of r^2 and thereby calculate the power recorded in the figure. In these calculations, the susceptibility allele was assumed to have a frequency of 10%, a GRR of 1.3 and a multiplicative model. Actual samples sizes in which the test variant is typed are listed in the legend. Power was calculated using the on-line genetic power calculator (Purcell *et al*., 2003).

Genome-wide association studies (GWAS)

One logical way to improve the odds of identifying susceptibility factors would be to consider all common variation rather than just a single randomly selected or candidate variant. If all common variation were to be typed in a study then this study would be sure to include an analysis of the relevant variants. In this situation, concerns about prior odds might be ignored and tests simply interpreted after some correction for multiple testing. However, rather predictably, nothing is gained by adopting this approach to analysis since the statistical penalty required to correct for multiple testing is equivalent to that incurred by allowing for the prior odds (Freimer and Sabatti, 2004). This is not surprising since both are simply a reflection of the size of the genome. This type of comprehensive (direct) GWAS would seem to be ideal but in fact the extent of LD between common variants is so extensive that an indirect screen involving just a fraction of the markers and relying on LD between tested and untested variants, enables a large proportion (typically >80%) of common variation to be screened in a highly efficient manner (Pe'er *et al*., 2006). Direct GWAS remain beyond affordable and practical technologies at this time but indirect GWAS are possible and have proven to be a highly successful means to identify common variants influencing susceptibility to complex disease (Duerr *et al*., 2006; Easton *et al*., 2007; Gudmundsson *et al*., 2007; Hampe *et al*., 2007; Helgadottir *et al*., 2007; Hunter *et al*., 2007; Libioulle *et al*., 2007; McPherson *et al*., 2007; Rioux *et al*., 2007; Saxena *et al*., 2007; Scott *et al*., 2007; Sladek *et al*., 2007; Stacey *et al*., 2007; Steinthorsdottir *et al*., 2007; WTCCC, 2007; Yeager *et al*., 2007).

Testing for association indirectly introduces another variable which influences the power to identify relevant loci, the extent of LD between the tested variant and the causative allele (Moskvina and O'Donovan, 2007). Figure 7 shows the effects of LD on power to identify meaningful association in studies of differing size.

From Fig. 7 it is clear that power falls dramatically as LD declines unless study samples are large enough as to ensure 'reserve power'. In a study involving 10 000 cases and 10 000 controls there would be little difference in power between causative variants and those in LD with $r^2 > 0.8$ (when considering variants with a frequency of 10% that increase risk by a factor of 1.3 or more). In the context of the MHC, even lower levels of LD can generate highly significant associations at test markers as the signal from the causative (DRB1*1501) allele is so strong.

One of the most notable features of the GWAS completed to date is the consistent observation that association tests are modestly inflated at neutral markers (i.e. those that do not influence susceptibility) in comparison with what would be expected if sampling error were the only source of variance. Exploring this systematic

'genomic inflation' Clayton *et al.* (2005) established that this modest but discernable effect stems from two influences, population stratification and differential 'missingness'. Population stratification refers to the generation of a case-control allele frequency difference due to systematic difference in ancestry between the cases and controls, in other words incomplete matching of cases and controls with regard to ancestry (Thomas and Witte, 2002). Although this effect has long been suggested as a source of false-positive association (Lander and Schork, 1994), the results from GWAS thus far published have empirically confirmed the prediction that the effect would rarely account for anything more than modest inflation of association (Cardon and Bell, 2001). In addition the WTCCC has shown that with very few exceptions allele frequencies do not vary significantly across the UK thereby confirming that within populations like the UK hidden stratification will rarely if ever produce more than modest inflation in the evidence for association (WTCCC, 2007). Unfortunately the MHC is one of the loci where allele frequencies do vary considerable across the country thereby raising the possibility that population stratification could confound the analysis of this locus if cases and controls are not adequately matched. However, having completed a GWAS it is straightforward to identify ancestry and compensate for any stratification (Devlin and Roeder, 1999; Bacanu *et al.*, 2000; Devlin *et al.*, 2004). By studying individuals whose ancestry is known to involve a mix of ethnic groups, which vary in their susceptibility to multiple sclerosis, population stratification can actually be used to help map risk loci (Smith *et al.*, 2004). Employing this admixture approach in African American patients Reich *et al.* (2005) identified a region on chromosome 1 where European ancestry was in statistically significant excess but this group has not yet been able to fine map the region and identify the relevant gene. Interestingly these researchers found no evidence for any distortion in ancestry in the MHC suggesting that the DRB1*1503 allele which is common in African individuals likely confers the same risk as the DRB1*1501 allele, which is more common in Europeans. In short these data suggest that the difference in risk seen between African and European individuals is unlikely to stem from the MHC and may well be determined by the yet to be defined locus on chromosome 1.

Differential missingness refers to the allele frequency difference that develops between cases and controls when genotyping failure is non-random with respect to genotype and differs in extent between cases and controls, i.e. when there is a difference in the amount of non-random missing information between cases and controls (Clayton *et al.*, 2005). In fact, genotyping failure is almost always non-random with respect to genotype with the result that genotyping efficiency, the extent to which genotyping is complete, is one of the most valuable measure of data quality. Only analysing markers with adequate levels of genotyping efficiency and no significant difference in genotyping efficiency between the cases and controls minimises the effects of this phenomenon. Since the perturbing influences of this effect are dependant upon MAF the genotyping efficiency threshold required must be more stringent for markers with MAF of <10%.

Recent progress

Using genomic convergence Fernald *et al.* (2005) identified a short list of multiple sclerosis candidate genes showing the greatest support for relevance in the existing literature. Prominent amongst these was the interleukin-7 receptor (IL7R) (Fernald *et al.*, 2005), a gene that had previously been identified as a candidate and studied by groups from Australia (Teutsch *et al.*, 2003) and Sweden (Zhang *et al.*, 2005). Following on from their analysis of genomic convergence Gregory *et al.* (2007) established significant association with the IL7R SNP rs6897932 ($P = 2.9 \times 10^{-7}$); simultaneously the Swedish group reported their extended analysis of IL7R and thereby replicated the association with rs6897932 (Lundmark *et al.*, 2007). This SNP codes for a non-synonymous variation in the alternatively spliced exon 6 of the IL7R gene. In a functional assay, Gregory *et al.* (2007) have also shown that the multiple sclerosis associated allele of rs6897932 increases the proportion of gene product in which exon 6 is skipped and for which the receptor is therefore soluble as opposed to membrane-bound. These observations predict that IL7 signalling should be impaired in multiple sclerosis, an observation which has been independently confirmed (Cox *et al.*, 2005). Quite why impaired IL7R signalling increases the risk of developing multiple sclerosis remains unknown, and of course it is always possible that the effect of this variant on IL7 signalling is an epiphenomenon with some other as yet untested function of this gene being more relevant to susceptibility. IL7R is thus the first confirmed non-MHC association in multiple sclerosis.

Alongside these candidate gene efforts, 2007 saw the publication of two GWAS studies in multiple sclerosis. In the first, the IMSGC screened 931 trio families (half from the US and half from the UK) using 334 923 SNPs (IMSGC, 2007). As would be predicted the limited power provided by 931 trio families meant that no unequivocal associations were identified in the screening phase, outside of the expected signals from the MHC. However, by utilizing additional controls from the WTCCC ($n = 1475$) and the National Institutes of Mental Health ($n = 956$) along with candidate gene information, a short list of 110 loci were followed up in an additional 2931 cases and 4205 controls. In the final analysis (employing a total of 12 360 individuals), association with rs6897932 (the IL7R associated SNP) was confirmed and significant association was also established with rs12722489 ($P = 3.0 \times 10^{-8}$) and rs2104286 ($P = 2.2 \times 10^{-7}$) from the interleukin-2 receptor (IL2R) gene making this the second non-MHC locus to be established in multiple sclerosis (IMSGC, 2007). In the second GWAS, performed by the WTCCC (Burton *et al.*, 2007),

975 cases and 1466 controls were screened with 12 374 non-synonymous SNPs (nsSNPs). Again, the limited power provided by the cohort size meant that the screen failed to identify any unequivocally associated markers. However, it is relevant to note that rs6897932 was the eighth most associated marker identified, confirming that a GWAS-based approach would have identified this association had it not already been established through the candidate gene approach.

Attempts to follow up on the other potential associations identified in these screens are underway alongside additional screens which will help to refine the ranking of tested variants.

Putting things in context

It is worth pausing to consider the nature of these new findings. Taking the IL7R association as an example we can see that the multiple sclerosis associated allele of rs6897932 has a frequency of 72% which means that ~9 out of every 10 white Europeans carry this risk allele, which therefore certainly would qualify as a common variant. The allele is estimated to increase the risk of the disease by a factor of just 1.2. Using these parameters we can calculate the significance level (*P*-value) that would be expected in attempts to replicate this finding as shown in Fig. 8.

It is clear from this figure that a replication study will need to involve at least 2000 cases and 2000 controls if it is to have >95% power to demonstrate a nominally significant *P*-value of 5%. Most attempts at replication involving more than 600 cases and 600 controls will be expected to yield a *P*-value of <5% but not all. Studies with less than 600 cases and 600 controls are unlikely to identify even nominally significant association. It will be important to keep these values in mind when interpreting replication studies. If a study involving just 400 cases and 400 controls fails to identify nominally significant association this should not be interpreted as evidence that rs6897932 is not relevant in the tested population. This is perhaps the least likely explanation.

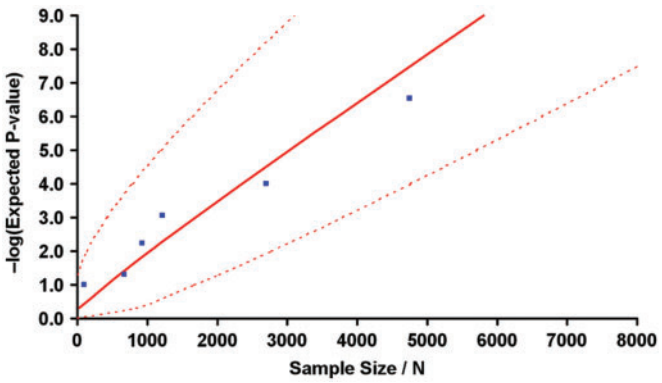

Fig. 8 Expected *P*-value in follow-up studies of rs6897932, the IL7R-associated SNP. The red line indicates the expected *P*-value and the dotted lines the 95% confidence intervals on this estimate (plotted as the negative log). It can thus be expected that 95% of the time the observed *P*-value with fall in this space. The blue dots indicate the studies already reported concerning this locus (the first two studies did not consider this variant directly but it is expected that the observed signal was due to LD with rs6897932). From left to right the studies are Teutsch (2003), Zhang (2005), IMSGC (screening phase) (2007), WTCCC (2007), Lundmark (2007) and Gregory (2007). Expected *P*-values and confidence intervals were calculated using the on-line genetic power calculator (Purcell *et al.*, 2003).

Taking these estimates of effect size and allele frequency we can calculate the lod score that rs6897932 would be expected to generate in a set of 100 sib pairs. This turns out to be <0.01! In short, loci such as rs6897932 will not be expected to generate any linkage signals discernable in previously published linkage screens. Thus, any apparent concordance between identified susceptibility loci and previously reported linkage peaks is entirely coincidental.

In conclusion

An analogy might serve to summarize the relative strengths of the many and varied methods which have been used to try and unravel the complex genetics underlying susceptibility to multiple sclerosis. Epidemiological analysis might be likened to a hand-held magnifying glass; it has allowed us to demonstrate that there are genetic factors to be found but because of its inherent imprecision and vulnerability to confounding is unable to reveal any greater detail. Linkage analysis on the other hand can be likened to a light microscope, large details such as the relevance of the MHC can be seen but this approach lacks the resolution needed to identify any other detail. It would be wholly inaccurate to infer that the failure of this insensitive instrument to identify any other detail implies that there are no further genes involved or that additional genes will be of no biological importance. GWAS provides us with the equivalent of an electron microscope, using this tool we are finally starting to identify relevant non-MHC loci and unravel the nature of susceptibility to multiple sclerosis. If funding agencies can be persuaded to follow this long road to its logical conclusion and support a 10 000-patient strong GWAS along with the necessary replication and fine-mapping efforts we can expect that most of the relevant common variation could be defined. Quite what the genetic landscape of multiple sclerosis will look like is hard to predict. It remains possible that any one of the ultimately identified loci could be especially informative about the nature of the disease or that the modest marginal effects attributable to individual loci might interact to produce larger effects, with subsets of risk loci implicating particularly important pathways. There is not much evidence for this in the data available to date and in other diseases it seems that effects are largely independent and additive (Weedon *et al.*, 2006), in which case we are only likely to be able to interpret these data and make use of them to define the immunological (or other) deficits which are responsible for susceptibility to multiple sclerosis once the catalogue of the relevant variants is comprehensive. 'This is not the end, it is not even the beginning of the end. But it is, perhaps, the end of the beginning' - Churchill 1942.

Acknowledgements

I would like to thank all my colleagues in the International Multiple Sclerosis Genetics Consortium (IMSGC), the Genetic Analysis of Multiple sclerosis in EuropeanS (GAMES) collaborative group and the Wellcome Trust Case Control Consortium (WTCCC) for all their support and tireless efforts to move the genetics of multiple sclerosis forward. I would especially like to thank An Goris for her careful scrutiny of the manuscript and her helpful comments. Funding to pay the Open Access charges for this paper was provided by the Wellcome Trust (grant ref. 084702).

References

Akesson E, Oturai A, Berg J, Fredrikson S, Andersen O, Harbo HF, et al. A genome-wide screen for linkage in Nordic sib-pairs with multiple sclerosis. Genes Immun 2002; 3: 279–85.

Allcock RJ, Atrazhev AM, Beck S, de Jong PJ, Elliott JF, Forbes S, et al. The MHC haplotype project: a resource for HLA-linked association studies. Tissue Antigens 2002; 59: 520–1.

Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 2000; 66: 1933–44.

Ban M, Stewart GJ, Bennetts BH, Heard R, Simmons R, Maranian M, et al. A genome screen for linkage in Australian sibling-pairs with multiple sclerosis. Genes Immun 2002; 3: 464–9.

Barcellos LF, Sawcer S, Ramsay PP, Baranzini SE, Thomson G, Briggs F, et al. Heterogeneity at the HLA-DRB1 locus and risk for multiple sclerosis. Hum Mol Genet 2006; 15: 2813–24.

Broadley S, Sawcer S, D'Alfonso S, Hensiek A, Coraddu F, Gray J, et al. A genome screen for multiple sclerosis in Italian families. Genes Immun 2001; 2: 205–10.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 2007; 39: 1329–37.

Burwick RM, Ramsay PP, Haines JL, Hauser SL, Oksenberg JR, Pericak-Vance MA, et al. APOE epsilon variation in multiple sclerosis susceptibility and disease severity: some answers. Neurology 2006; 66: 1373–83.

Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001; 2: 91–9.

Carton H, Vlietinck R, Debruyne J, De Keyser J, D'Hooghe MB, Loos R, et al. Risks of multiple sclerosis in relatives of patients in Flanders, Belgium. J Neurol Neurosurg Psychiatry 1997; 62: 329–33.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 2005; 37: 1243–6.

Cohen D, Cohen O, Marcadet A, Massart C, Lathrop M, Deschamps I, et al. Class II HLA-DC beta-chain DNA restriction fragments differentiate among HLA-DR2 individuals in insulin-dependent diabetes and multiple sclerosis. Proc Natl Acad Sci USA 1984; 81: 1774–8.

Compston A. Making progress on the natural history of multiple sclerosis. Brain 2006; 129: 561–3.

Compston A, Coles A. Multiple sclerosis. Lancet 2002; 359: 1221–31.

Compston A, Confavreux C, Lassmann H, McDonald I, Miller D, Noseworthy J, et al. McAlpine's multiple sclerosis. London: Churchill Livingstone; 2006; 113–81.

Compston DA, Batchelor JR, McDonald WI. B-lymphocyte alloantigens associated with multiple sclerosis. Lancet 1976; 2: 1261–5.

Confavreux C, Vukusic S. Age at disability milestones in multiple sclerosis. Brain 2006a; 129: 595–605.

Confavreux C, Vukusic S. Natural history of multiple sclerosis: a unifying concept. Brain 2006b; 129: 606–16.

Coraddu F, Sawcer S, D'Alfonso S, Lai M, Hensiek A, Solla E, et al. A genome screen for multiple sclerosis in Sardinian multiplex families. Eur J Hum Genet 2001; 9: 621–6.

Cox AL, Thompson SA, Jones JL, Robertson VH, Hale G, Waldmann H, et al. Lymphocyte homeostasis following therapeutic lymphocyte depletion in multiple sclerosis. Eur J Immunol 2005; 35: 3332–42.

de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 2006; 38: 1166–72.

de Jong BA, Huizinga TW, Zanelli E, Giphart MJ, Bollen EL, Uitdehaag BM, et al. Evidence for additional genetic risk indicators of relapse-onset MS within the HLA region. Neurology 2002; 59: 549–55.

Devlin B, Bacanu SA, Roeder K. Genomic control to the extreme. Nat Genet 2004; 36: 1129–30; author reply 1131.

Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999; 55: 997–1004.

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 2006; 314: 1461–3.

Dyment DA, Cader MZ, Willer CJ, Risch N, Sadovnick AD, Ebers GC. A multigenerational family with multiple sclerosis. Brain 2002; 125: 1474–82.

Dyment DA, Herrera BM, Cader MZ, Willer CJ, Lincoln MR, Sadovnick AD, et al. Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance. Hum Mol Genet 2005; 14: 2019–26.

Dyment DA, Sadovnick AD, Willer CJ, Armstrong H, Cader ZM, Wiltshire S, et al. An extended genome scan in 442 Canadian multiple sclerosis-affected sibships: a report from the Canadian Collaborative Study Group. Hum Mol Genet 2004; 13: 1005–15.

Dyment DA, Yee IM, Ebers GC, Sadovnick AD. Multiple sclerosis in stepsiblings: recurrence risk and ascertainment. J Neurol Neurosurg Psychiatry 2006; 77: 258–9.

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007; 447: 1087–93.

Ebers GC, Kukay K, Bulman DE, Sadovnick AD, Rice G, Anderson C, et al. A full genome search in multiple sclerosis. Nat Genet 1996; 13: 472–6.

Ebers GC, Sadovnick AD, Dyment DA, Yee IM, Willer CJ, Risch N. Parent-of-origin effect in multiple sclerosis: observations in half-siblings. Lancet 2004; 363: 1773–4.

Ebers GC, Sadovnick AD, Risch NJ. A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. Nature 1995; 377: 150–1.

Ebers GC, Yee IM, Sadovnick AD, Duquette P. Conjugal multiple sclerosis: population-based prevalence and recurrence risks in offspring. Canadian Collaborative Study Group. Ann Neurol 2000; 48: 927–31.

Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genet 2005; 6: 18.

Eraksoy M, Kurtuncu M, Akman-Demir G, Kilinc M, Gedizlioglu M, Mirza M, et al. A whole genome screen for linkage in Turkish multiple sclerosis. J Neuroimmunol 2003; 143: 17–24.

Fernald GH, Yeh RF, Hauser SL, Oksenberg JR, Baranzini SE. Mapping gene activity in complex disorders: Integration of expression and genomic scans for multiple sclerosis. J Neuroimmunol 2005; 167: 157–69.

Fogdell-Hahn A, Ligers A, Gronning M, Hillert J, Olerup O. Multiple sclerosis: a modifying influence of HLA class I genes in an HLA class II associated autoimmune disease. Tissue Antigens 2000; 55: 140–8.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449: 851–61.

Freimer N, Sabatti C. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet 2004; 36: 1045–51.

French Research Group on Multiple Sclerosis. Multiple sclerosis in 54 twinships: concordance rate is independent of zygosity. Ann Neurol 1992; 32: 724–7.

Genetic Analysis of Multiple Sclerosis in EuropeanS (GAMES) and the Transatlantic Multiple Sclerosis Genetics Cooperative. A meta-analysis

of whole genome linkage screens in multiple sclerosis. J Neuroimmunol 2003; 143: 39–46.
Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 2002; 54: 22–33.
Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 2001; 69: 1357–69.
Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, et al. Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet 2007; 39: 1083–91.
Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007; 39: 631–7.
Guo SW. Sibling recurrence risk ratio as a measure of genetic effect: caveat emptor! Am J Hum Genet 2002; 70: 818–9.
Haines JL, Ter-Minassian M, Bazyk A, Gusella JF, Kim DJ, Terwedow H, et al. A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. The Multiple Sclerosis Genetics Group. Nat Genet 1996; 13: 469–71.
Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007; 39: 207–11.
Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, Kyvik KO. Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. Mult Scler 2005; 11: 504–10.
Harbo HF, Lie BA, Sawcer S, Celius EG, Dai KZ, Oturai A, et al. Genes in the HLA class I region may contribute to the HLA class II-associated genetic susceptibility to multiple sclerosis. Tissue Antigens 2004; 63: 237–47.
Harding AE, Sweeney MG, Miller DH, Mumford CJ, Kellar-Wood H, Menard D, et al. Occurrence of a multiple sclerosis-like illness in women who have a Leber's hereditary optic neuropathy mitochondrial DNA mutation. Brain 1992; 115 (Pt 4): 979–89.
Hauser MA, Li YJ, Takeuchi S, Walters R, Noureddine M, Maready M, et al. Genomic convergence: identifying candidate genes for Parkinson's disease by combining serial analysis of gene expression and genetic linkage. Hum Mol Genet 2003; 12: 671–7.
Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 2007; 316: 1491–3.
Hensiek AE, Roxburgh R, Smilie B, Coraddu F, Akesson E, Holmans P, et al. Updated results of the United Kingdom linkage-based genome screen in multiple sclerosis. J Neuroimmunol 2003a; 143: 25–30.
Hensiek AE, Sawcer SJ, Compston DA. Searching for needles in haystacks-the genetics of multiple sclerosis and other common neurological diseases. Brain Res Bull 2003b; 61: 229–34.
Hensiek AE, Seaman SR, Barcellos LF, Oturai A, Eraksoi M, Cocco E, et al. Familial effects on the clinical course of multiple sclerosis. Neurology 2007; 68: 376–83.
Herrera BM, Ramagopalan SV, Orton S, Chao MJ, Yee IM, Sadovnick AD, et al. Parental transmission of MS in a population-based Canadian cohort. Neurology 2007; 69: 1208–12.
Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002; 4: 45–61.
Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nat Rev Genet 2004; 5: 889–99.
Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, et al. Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. Immunogenetics 2008; 60: 1–18.
Human genome project. Finishing the euchromatic sequence of the human genome. Nature 2004; 431: 931–45.
Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007; 39: 870–4.
Hupperts R, Broadley S, Mander A, Clayton D, Compston DA, Robertson NP. Patterns of disease in concordant parent-child pairs with multiple sclerosis. Neurology 2001; 57: 290–5.
International HapMap Project. Nature 2003; 426: 789–96.
International Multiple Sclerosis Genetics Consortium (IMSGC). Enhancing linkage analysis of complex disorders: an evaluation of high-density genotyping. Hum Mol Genet 2004; 13: 1943–9.
International Multiple Sclerosis Genetics Consortium (IMSGC). A high-density screen for linkage in multiple sclerosis. Am J Hum Genet 2005; 77: 454–67.
International Multiple Sclerosis Genetics Consortium (IMSGC). Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. N Engl J Med 2007; 357: 851–62.
Ioannidis JP. Genetic associations: false or true? Trends Mol Med 2003; 9: 135–8.
Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol 2006; 164: 609–14.
Islam T, Gauderman WJ, Cozen W, Hamilton AS, Burnett ME, Mack TM. Differential twin concordance for multiple sclerosis by latitude of birthplace. Ann Neurol 2006; 60: 56–64.
Jersild C, Fog T, Hansen GS, Thomsen M, Svejgaard A, Dupont B. Histocompatibility determinants in multiple sclerosis, with special reference to clinical course. Lancet 1973; 2: 1221–5.
Jersild C, Svejgaard A, Fog T. HL-A antigens and multiple sclerosis. Lancet 1972; 1: 1240–1.
Kantarci OH, Barcellos LF, Atkinson EJ, Ramsay PP, Lincoln R, Achenbach SJ, et al. Men transmit MS more often to their children vs women: the Carter effect. Neurology 2006; 67: 305–10.
Kenealy SJ, Babron MC, Bradford Y, Schnetz-Boutaud N, Haines JL, Rimmler JB, et al. A second-generation genomic screen for multiple sclerosis. Am J Hum Genet 2004; 75: 1070–8.
Koeleman BP, Herr MH, Zavattari P, Dudbridge F, March R, Campbell D, et al. Conditional ETDT analysis of the human leukocyte antigen region in type 1 diabetes. Ann Hum Genet 2000; 64: 215–21.
Kremenchutzky M, Rice GP, Baskerville J, Wingerchuk DM, Ebers GC. The natural history of multiple sclerosis: a geographically based study 9: observations on the progressive phase of the disease. Brain 2006; 129: 584–94.
Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001; 27: 234–6.
Kuokkanen S, Gschwend M, Rioux JD, Daly MJ, Terwilliger JD, Tienari PJ, et al. Genomewide scan of multiple sclerosis in Finnish multiplex families. Am J Hum Genet 1997; 61: 1379–87.
Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265: 2037–48.
Lennon VA, Kryzer TJ, Pittock SJ, Verkman AS, Hinson SR. IgG marker of optic-spinal multiple sclerosis binds to the aquaporin-4 water channel. J Exp Med 2005; 202: 473–7.
Lennon VA, Wingerchuk DM, Kryzer TJ, Pittock SJ, Lucchinetti CF, Fujihara K, et al. A serum autoantibody marker of neuromyelitis optica: distinction from multiple sclerosis. Lancet 2004; 364: 2106–12.
Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 2007; 3: e58.
Ligers A, Dyment DA, Willer CJ, Sadovnick AD, Ebers G, Risch N, et al. Evidence of linkage with HLA-DR in DRB1*15-negative families with multiple sclerosis. Am J Hum Genet 2001; 69: 900–3.
Lincoln MR, Montpetit A, Cader MZ, Saarela J, Dyment DA, Tiislar M, et al. A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. Nat Genet 2005; 37: 1108–12.
Lindsey JW. Familial recurrence rates and genetic models of multiple sclerosis. Am J Med Genet A 2005; 135: 53–8.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003; 33: 177–82.

Lundmark F, Duvefelt K, Iacobaeus E, Kockum I, Wallstrom E, Khademi M, et al. Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. Nat Genet 2007; 39: 1108–13.

Marrosu MG, Murru MR, Costa G, Murru R, Muntoni F, Cucca F. DRB1-DQA1-DQB1 loci and multiple sclerosis predisposition in the Sardinian population. Hum Mol Genet 1998; 7: 1235–7.

Marrosu MG, Murru R, Murru MR, Costa G, Zavattari P, Whalen M, et al. Dissection of the HLA association with multiple sclerosis in the founder isolated population of Sardinia. Hum Mol Genet 2001; 10: 2907–16.

McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, et al. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. Ann Neurol 2001; 50: 121–7.

McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, et al. A common allele on chromosome 9 associated with coronary heart disease. Science 2007; 316: 1488–91.

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, et al. A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 2005; 76: 634–46.

Modin H, Masterman T, Thorlacius T, Stefansson M, Jonasdottir A, Stefansson K, et al. Genome-wide linkage screen of a consanguineous multiple sclerosis kinship. Mult Scler 2003; 9: 128–34.

Moskvina V, O'Donovan MC. Detailed analysis of the relative power of direct and indirect association studies and the implications for their interpretation. Hum Hered 2007; 64: 63–73.

Mumford CJ, Wood NW, Kellar-Wood H, Thorpe JW, Miller DH, Compston DA. The British Isles survey of multiple sclerosis in twins. Neurology 1994; 44: 11–5.

Naito S, Namerow N, Mickey MR, Terasaki PI. Multiple sclerosis: association with HL-A3. Tissue Antigens 1972; 2: 1–4.

Oksenberg JR, Barcellos LF, Cree BA, Baranzini SE, Bugawan TL, Khan O, et al. Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans. Am J Hum Genet 2004; 74: 160–7.

Olerup O, Hillert J. HLA class II-associated genetic susceptibility in multiple sclerosis: a critical evaluation. Tissue Antigens 1991; 38: 1–15.

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 2006; 38: 663–7.

Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 1997; 276: 2045–7.

Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant . . . or not? Hum Mol Genet 2002; 11: 2417–23.

Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003; 19: 149–50.

Ramagopalan SV, Morris AP, Dyment DA, Herrera BM, DeLuca GC, Lincoln MR, et al. The inheritance of resistance alleles in multiple sclerosis. PLoS Genet 2007; 3: 1607–13.

Reich D, Patterson N, Jager PL, McDonald GJ, Waliszewska A, Tandon A, et al. A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. Nat Genet 2005; 37: 1113–8.

Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001; 17: 502–10.

Riordan-Eva P, Sanders MD, Govan GG, Sweeney MG, Da Costa J, Harding AE. The clinical features of Leber's hereditary optic neuropathy defined by the presence of a pathogenic mitochondrial DNA mutation. Brain 1995; 118 (Pt 2): 319–37.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007; 39: 596–604.

Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 1990; 46: 222–8.

Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996; 273: 1516–7.

Ristori G, Cannoni S, Stazi MA, Vanacore N, Cotichini R, Alfo M, et al. Multiple sclerosis in twins from continental Italy and Sardinia: a nationwide study. Ann Neurol 2006; 59: 27–34.

Robertson NP, Fraser M, Deans J, Clayton D, Walker N, Compston DA. Age-adjusted recurrence risks for relatives of patients with multiple sclerosis. Brain 1996; 119 (Pt 2): 449–55.

Robertson NP, O'Riordan JI, Chataway J, Kingsley DP, Miller DH, Clayton D, et al. Offspring recurrence rates and clinical characteristics of conjugal multiple sclerosis. Lancet 1997; 349: 1587–90.

Rubio JP, Bahlo M, Butzkueven H, van Der Mei IA, Sale MM, Dickinson JL, et al. Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. Am J Hum Genet 2002; 70: 1125–37.

Sadovnick AD, Baird PA, Ward RH. Multiple sclerosis: updated risks for relatives. Am J Med Genet 1988; 29: 533–41.

Sawcer S. A new era in the genetic analysis of multiple sclerosis. Curr Opin Neurol 2006; 19: 237–41.

Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, et al. A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. Nat Genet 1996; 13: 464–8.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007; 316: 1331–6.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007; 316: 1341–5.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007; 445: 881–5.

Smith DJ, Lusis AJ. The allelic structure of common disease. Hum Mol Genet 2002; 11: 2455–61.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, et al. A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 2004; 74: 1001–13.

Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007; 39: 865–9.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 2007; 39: 770–5.

Tabor HK, Risch NJ, Myers RM. Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 2002; 3: 391–7.

Terasaki PI, Park MS, Opelz G, Ting A. Multiple sclerosis and high incidence of a B lymphocyte antigen. Science 1976; 193: 1245–7.

Teutsch SM, Booth DR, Bennetts BH, Heard RN, Stewart GJ. Identification of 11 novel and common single nucleotide polymorphisms in the interleukin-7 receptor-alpha gene and their associations with multiple sclerosis. Eur J Hum Genet 2003; 11: 509–15.

Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002; 11: 505–12.

Vartdal F, Sollid LM, Vandvik B, Markussen G, Thorsby E. Patients with multiple sclerosis carry DQB1 genes which encode shared polymorphic amino acid sequences. Hum Immunol 1989; 25: 103–10.

Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 2004; 96: 434–42.

Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 2003; 4: 587–97.

Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005; 6: 109–18.

Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med 2006; 3: e374.

Wellcome Trust Case Control Consortium (WTCC). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; 447: 661–78.

Willer CJ, Dyment DA, Cherny S, Ramagopalan SV, Herrera BM, Morrison KM, et al. A genome-wide scan in forty large pedigrees with multiple sclerosis. J Hum Genet 2007; 52: 955–62.

Willer CJ, Dyment DA, Risch NJ, Sadovnick AD, Ebers GC. Twin concordance and sibling recurrence rates in multiple sclerosis. Proc Natl Acad Sci USA 2003; 100: 12877–82.

Wingerchuk DM, Lennon VA, Pittock SJ, Lucchinetti CF, Weinshenker BG. Revised diagnostic criteria for neuromyelitis optica. Neurology 2006; 66: 1485–9.

Yang Q, Khoury MJ, Friedman J, Little J, Flanders WD. How many genes underlie the occurrence of common complex diseases in the population? Int J Epidemiol 2005; 34: 1129–37.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007; 39: 645–9.

Yeo TW, De Jager PL, Gregory SG, Barcellos LF, Walton A, Goris A, et al. A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol 2007; 61: 228–36.

Zhang Z, Duvefelt K, Svensson F, Masterman T, Jonasdottir G, Salter H, et al. Two genes encoding immune-regulatory molecules (LAG3 and IL7R) confer susceptibility to multiple sclerosis. Genes Immun 2005; 6: 145–52.