REVIEW

Susceptibility gene discovery for common metabolic and endocrine traits

M I McCarthy

Imperial College Genetics and Genomics Research Institute and Division of Medicine, Imperial College, London, UK

(Requests for offprints should be addressed to M I McCarthy, Imperial College Genetics and Genomics Research Institute, Imperial College (Hammersmith Campus), London W12 0NN, UK; Email: m.mccarthy@ic.ac.uk)

Abstract

Almost all major causes of ill-health and premature death in human societies worldwide – including cancer, cardiovascular disease, diabetes and many infectious diseases – are, at least in part, genetically determined. Typically, risk of succumbing to one of these illnesses is thought to depend on both the individual repertoire of variation within a number of key susceptibility genes and the history of exposure to relevant environmental factors. For many of these conditions, the molecular basis of disease pathogenesis remains obscure. This represents a major obstacle to development of improved, rational strategies for disease treatment, prevention and eradication. It is easy therefore to appreciate the importance attached to efforts to deliver more comprehensive understanding of the molecular basis of disease pathogenesis. Nor is it hard to understand that identification of major susceptibility genes should highlight those components of molecular machinery that are critical for the preservation of normal health.

The benefits promised are great, but progress to gene identification in multifactorial traits has been rather disappointing to date. Why is this? This review aims to answer this question by describing current and future approaches to gene discovery in multifactorial traits. The examples quoted will mostly relate to type 2 diabetes, but the issues and approaches are generic, and apply equally to other multifactorial traits in the endocrine and metabolic arena – type 1 diabetes; obesity; hyperlipidaemia; autoimmune thyroid disease; polycystic ovarian syndrome – and beyond.

Journal of Molecular Endocrinology (2002) **28,** 1–17

The challenge of gene discovery in multifactorial diseases

In the past decade, gene identification for monogenic (or Mendelian) diseases has become an increasingly routine affair. Causative variants for several hundred different single-gene disorders have been pinpointed (Peltonen & McKusick 2001), and these have, in many cases, provided profound insights into fundamental biological processes. However, single-gene disorders are, by their very nature, relatively rare, and whilst the impact of the genomic variants responsible may be severe for those individuals (and families) affected, collectively they account for only a small proportion of illness within the population.

In contrast, we can expect the genetic variants that influence susceptibility to the dominant causes of morbidity and mortality in societies – for example, cardiovascular disease, diabetes, cancer – to have more modest effects at the individual level, but to have substantial impact within populations (Lander & Schork 1994, Vyse & Todd 1996). This distinction goes to the heart of the challenge presented by multifactorial traits. The variants that need to be identified are likely: to be common; to be present in both affected and unaffected individuals; to be associated with relatively modest increases in individual risk; to have subtle rather than disastrous effects on gene product function (e.g. via alterations in transcriptional regulation); and to interact in

0952–5041/02/028–001

complex, non-linear ways with other susceptibility factors contributing to disease (both genetic and environmental).

One way of looking at this distinction between monogenic and multifactorial traits is in terms of the correspondence between genotype and phenotype. Characteristically, for monogenic traits, this correspondence is close to 1:1. Thus, all individuals with cystic fibrosis have defective function of the cystic fibrosis transmembrane conductance regulator protein (CFTR) due to mutations in both copies of the *CFTR* gene, and all individuals with two severe mutations in *CFTR* inevitably develop cystic fibrosis (although genetic and environmental modifiers can vary the precise phenotypic expression) (Kiesewetter *et al.* 1993). In complex traits, this genotype–phenotype correspondence is much less tight. For one thing, the same phenotype may arise as a result of abnormalities in any one of (or combination of) several genes ('genetic heterogeneity'), or even, in certain circumstances from environmental exposures alone ('phenocopies'). For another, variation at any given site will not provide precise prediction of an individual's disease status ('incomplete penetrance'). A given variant may increase the individual risk of a given disease phenotype, but even this risk may be heavily dependent on the genetic and environmental context ('gene–gene' and 'gene–environment interaction').

If this were not demanding enough, multifactorial traits present additional complexities. First, there are often difficulties with diagnostic classification (e.g. what glucose level constitutes 'diabetes' (World Health Organisation Study Group 1985)?). How can one differentiate late-onset type 1 diabetes (T1D) from type 2 diabetes (T2D) (Tuomi *et al.* 1993)? Secondly, ascertainment of the family material, which is the basic substrate for most genetic research, may be problematical, especially in diseases of late-onset (Frayling *et al.* 1999). Thirdly, the assessment of the candidacy of particular genes and pathways is frustrated by ignorance of the biological basis of disease. (Is T2D primarily a disease of carbohydrate or lipid metabolism? Is the beta-cell, muscle, fat, liver or brain 'culpable' (Aitman *et al.* 1999)?) Fourthly, there may be marked ethnic heterogeneity – if Neel's 'thrifty genotype' explanation for the high prevalence of diabetes and obesity is correct (Neel 1982), it may well be that different ethnic groups have developed diverse molecular mechanisms to provide the metabolic efficiency that maximises survival during periods of erratic food supply (but which predisposes to obesity and diabetes in times of plenty).

Thus, whilst many of the tools employed in the dissection of complex traits are similar to those developed for, and successfully implemented in, studies of monogenic traits, there are necessarily substantial differences, both qualitative and quantitative, in the strategies adopted.

Tools of the trade

The analytical tools of the complex-trait mapper are based around the detection of signals for linkage and linkage disequilibrium (LD), and it is worth trying to disentangle these two related but distinct concepts.

Linkage

The independent segregation of chromosomes during meiosis ensures that alleles at two genes on different chromosomes are distributed randomly to gametes (the genes are 'unlinked'). However, when two genes lie on the same chromosome, their relationship following chromosomal segregation is determined by recombination between homologous chromosomes occurring during meiosis. The closer the physical location of the two genes, the less likely it is that a recombination event will separate them, and the more likely it is that alleles at those genes will be observed to co-segregate (into gametes, and thereby into offspring). This genetic 'linkage' provides a powerful tool for disease gene localisation (Ott 1999). All one needs, in principle, is a sufficiently large collection of families segregating the disease of interest (and hence assumed to be segregating the susceptibility genes for which one is searching), and a set of polymorphic markers, at known chromosomal locations, which can be typed to reveal patterns of chromosomal segregation in those pedigrees. Linkage analysis represents the computational tool which allows identification of those genomic regions which show statistically significant co-segregation with disease and are therefore likely to be harbouring susceptibility loci. Whilst ready access to large pedigrees, and a simple, defined genetic architecture, have made

linkage analysis the central tool for gene localisation in monogenic traits, precisely the same approaches are applicable, with some modification, to complex traits. The principal modifications include: (i) an emphasis on analysis of large numbers of small (nuclear) families or sibships rather than small numbers of large pedigrees (Davies *et al.* 1994, Lander & Schork 1994); (ii) use of 'non-parametric' model-free methods which do not require explicit description of the genetic architecture of the trait (Kruglyak *et al.* 1996), something scarcely possible for multifactorial diseases (Ott 1990); and (iii) ability to capitalise on underlying disease-related quantitative trait data (e.g. measures of insulin sensitivity for T2D) to complement analysis of dichotomous disease traits and, in many circumstances, offer increased power (Ghosh & Schork 1996, Almasy & Blangero 1998). Even so, as described below, the modest relative risks expected of most complex trait susceptibility loci set real limits to the reliable and robust detection of the linkage signals they may produce.

Linkage disequilibrium

Whilst linkage analysis looks at the effects of recombination events on the segregation of genes within families, LD analysis deals with the patterns of alleles within populations. LD is a special case of 'allelic association', that is characterised by the co-occurrence on a given chromosome of two alleles (from different loci) at a frequency different from that expected from the product of their individual frequencies (Lander & Schork 1994). For example, in European populations, the T allele at the −23 *HphI* polymorphism within the insulin gene, and the cluster of so-called class III alleles at the nearby *INS-VNTR* minisatellite are each present on about 30% of chromosomes (Bennett & Todd 1996). However, because of tight LD in the region, the frequency with which chromosomes carry both alleles is also ∼30% rather than the 9%, the product of their individual allele frequencies, that one would expect if they were in equilibrium (Fig. 1). The cardinal feature of LD, as opposed to simple allelic association, is that the two loci concerned are linked. Other mechanisms, such as latent population substructure, which can lead to associations between alleles at unlinked loci (i.e. association without LD), are troublesome from a methodological point of view, but are not generally of any great intrinsic biological interest (Lander & Schork 1994, Spielman & Ewens 1996).

To appreciate the ways in which linkage and LD analyses are deployed in the hunt for complex trait genes, it is important to understand a little more about the processes governing the development and dissipation of LD. LD around an allele arises, in the first place, either through natural selection or as a result of events and processes modifying the genetic composition of a population during its history; these include periods of small population size ('bottlenecks'), genetic admixture (due to interbreeding with a distinct population) and stochastic effects ('genetic drift'). At the same time, any LD established is gradually dissipated by the actions of recombination and mutation (Kruglyak 1999, Reich *et al.* 2001).

The most readily understood scenario for the generation of LD is provided by the 'founder effect'. Consider a modern population which arose through expansion of a small group of original founders (Finland or Tristan da Cunha are oft-quoted examples). Imagine that one of those founders carried a genetic variant that increases susceptibility to a given disease (but without a severe impact on reproductive potential). As that variant is passed down through subsequent generations, successive recombination events will mean that the descendent chromosomes on which that variant is carried become increasingly fragmented patchworks, reflecting those diverse parental and ancestral contributions. However, (very) close to the variant itself, the opportunities for such disruptive recombination events will have been limited, and many of the chromosomes carrying the variant will still resemble the original founding chromosome (and therefore each other). The consequence is a localised 'patch' where alleles on the ancestral chromosome are associated with (and in LD with) disease. Since chromosomes that carry the susceptibility variant will be over-represented amongst those with disease, it should be possible to find this patch by comparing disease cases and controls from the present-day population, and searching for regions where alleles show significant associations with disease.

Since it is underpinned by the same relationship between physical distance and recombination frequency, LD, just like linkage, can be used to localise susceptibility genes (Jorde 1995, 2000). However, because any residual LD will have

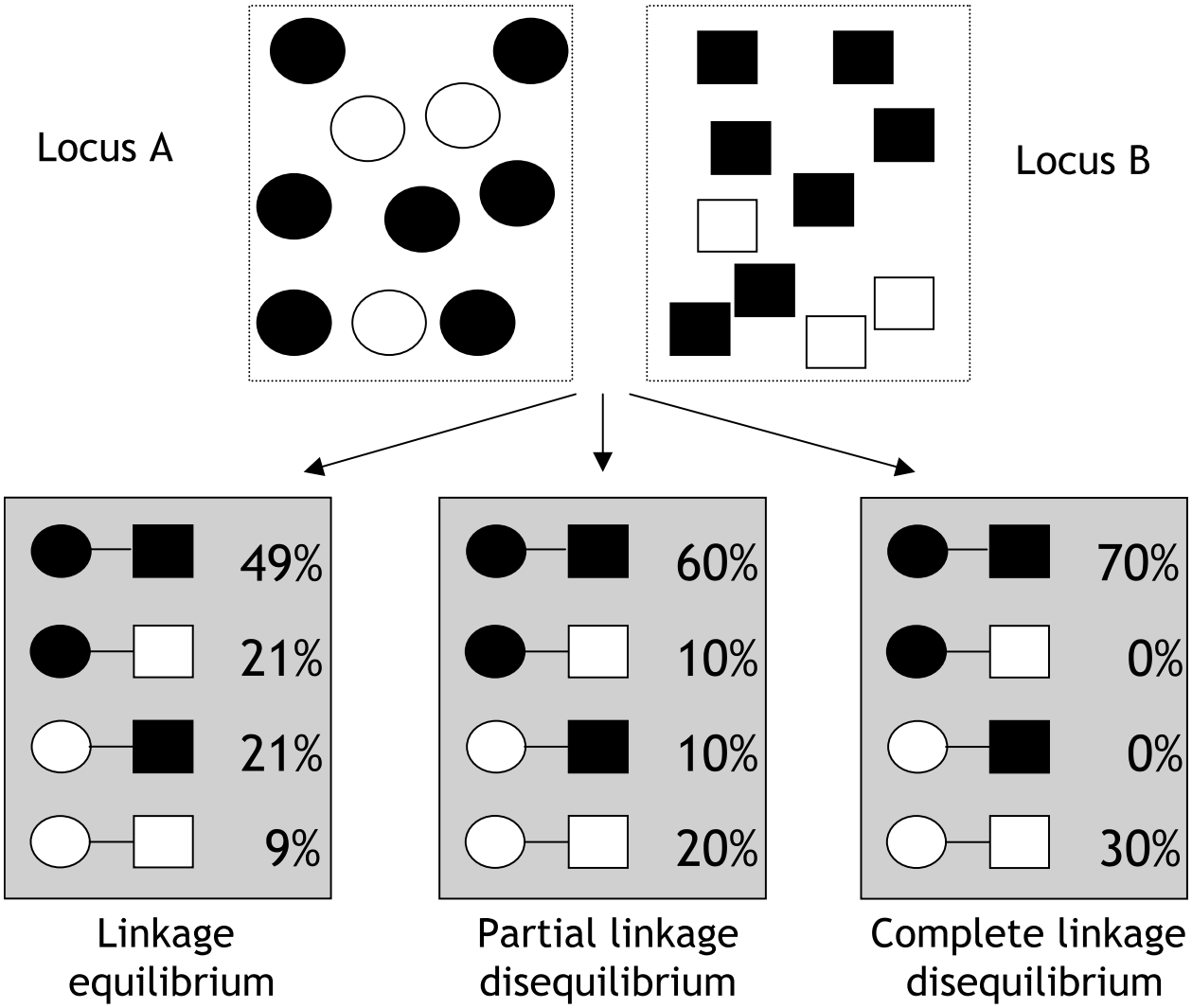

Figure 1 LD at a pair of polymorphic loci. The example shown is of two polymorphic variants, locus A (represented by circles) and locus B (squares). At each of these loci there are two alleles (represented by filled and open symbols); 'filled' alleles predominate (70%) at both loci. Assume that the loci are close neighbours on the same chromosome such that 'haplotypes' (comprising the alleles on a given chromosome) can be constructed. Clearly, there are four possible haplotypes, as shown in the lower part of the diagram. The relative proportions of those four haplotypes will be determined by the extent of LD between the two loci. If loci A and B are in linkage equilibrium (left panel), the haplotype frequencies are simply the products of the allele frequencies, and knowing an individual's genotype at locus A will provide no clues to their genotype at locus B. Alternatively, if (as in the other panels) loci A and B are in LD, haplotype frequencies will depart from this expectation, and certain haplotypes (right panel) may be absent or rare. Knowledge of an individual's genotype at locus A will provide some (middle panel) or complete (right panel) knowledge of the genotype at locus B.

'survived' the attrition of recombination events throughout a population's history, LD signals are less extensive than those arising from linkage studies (of the order of a few tens of kilobases, rather than a few tens of megabases) and, in principle, capable of providing much more precise localisation of disease genes (Lonjou *et al.* 1999, Abecasis *et al.* 2001, Reich *et al.* 2001). The main drawback in using LD to map genes is that the genomic extent, pattern and magnitude of LD in any given mapping situation (as defined by the combination of variant, disease and population

studied) is dependent on a range of highly variable and unpredictable factors, including population history, evolutionary selection, disease architecture and mutation rates (Risch 2000, Roses 2000, Weiss & Terwilliger 2000). Thus, the power and value of LD analysis in any given complex trait mapping effort is hard to gauge in advance. LD studies have certainly proven useful for disease gene localisation and identification in rare monogenic diseases in both population isolates and outbred populations (e.g. cystic fibrosis (Kerem *et al.* 1989)). The detection of association is the objective of most candidate gene studies for complex traits (Altshuler *et al.* 2000*a*), but in this case, the scale of the task is eased by the expectation that one might be detecting the actual aetiological variants (so that the degree of local LD is less of an issue). There are some promising examples of how LD can succeed in the more challenging task of localising complex trait genes within large genomic regions (Bennett & Todd 1996, Roses 2000, Hugot *et al.* 2001). However, substantial theoretical, methodological and practical obstacles remain to be overcome before one can become confident about the prospects for genome-wide analyses for LD (analogous to the genome-wide scans for linkage which are now routine) (Risch & Merikangas 1996, Weiss & Terwilliger 2000).

Strategies for gene discovery in multifactorial traits

Candidate gene studies

Conceptually, the simplest strategy for gene discovery in multifactorial traits is the 'candidate gene study' (Altshuler *et al.* 2000*a*, Cardon & Bell 2001). The usual procedure is to select a gene, usually on the basis of its known or presumed biological function, and the hypothesised relevance of that function to the disease of interest, and then to look for association between one or more variants in that gene and the disease phenotype. If a robust, statistically significant association is found, the implication is that the variant tested is either contributing directly to the phenotype or else is in LD with (and therefore relatively close to) such a variant. There are, of course, a number of alternative explanations including the possibility that the association is entirely spurious (due to type 1 error (Altshuler *et al.* 2000*a*)) or that the association has been the result of latent population stratification, through, for example, failure to match cases and control groups for ethnic background. This can result in non-linked genes appearing associated, i.e. association without LD (Williams *et al.* 1981, Lander & Schork 1994, Spielman & Ewens 1996). Family-based association methods (using parent–offspring trios or discordant sibling pairs) are a popular means of controlling for this second possibility, since their merit lies in generating a set of control chromosomes matched for parental origin to the disease chromosomes (Spielman & Ewens 1996, Boehnke & Langefeld 1998). The transmission disequilibrium test (TDT) (Spielman & Ewens 1996) has been the most widely applied of these family-based association tests, and in its simplest form involves measuring the frequency with which a given variant is transmitted from heterozygous parents to their offspring. Clearly, in normal circumstances, one would expect both alleles in a heterozygous parent to have an equal chance of being represented in their gametes, and subsequently in their offspring. Finding that a variant is significantly overtransmitted from heterozygous parents to affected offspring (in a set of parent–offspring trios ascertained for disease, for example) provides a simultaneous test of both association and linkage, which will not be deceived by association resulting from latent stratification. The TDT also provides an excellent tool for detecting parent-of-origin effects (Huxtable *et al.* 2000).

The T2D genetics literature is not unique in being populated by multiple association studies of candidate genes (McCarthy & Hitman 1993). Many positive associations have been reported but subsequent replication has proven the exception rather than the rule (Altshuler *et al.* 2000*a*). This confusing state of affairs is a consequence, in part, of the intrinsic 'biological' difficulties associated with complex trait genetics – the individual effect of any given variant is likely to be modest and to depend on genetic background and environmental exposure (Cardon & Bell 2001). However, this has undoubtedly been compounded by inadequacies in experimental method (small sample sizes; multiple testing leading to inflated type 1 error; publication bias; unrepresentative control populations).

Two examples amply illustrate the problem. Keavney *et al.* (2000) conducted a meta-analysis of published association studies relating *ACE* (the gene

for angiotensin-converting enzyme) I/D genotype to risk of myocardial infarction. Whilst the combined risk ratio for the 'at-risk' DD genotype in 35 published small studies (total of 3578 cases) was 1·57 (99% confidence interval: 1·38–1·78), the equivalent figures for 15 larger studies (11 492 cases) was 1·02 (0·95–1·11). Clearly, publication (and other) biases had produced a significant overestimation of effect in the smaller, less powerful studies.

Altshuler *et al.* (2000*a*) recently evaluated the *PPARG* Pro12 Ala variant in T2D. They convincingly demonstrated a modest but highly significant increase in relative risk (1·25, $P=0·002$) associated with the common Pro allele in analysis of several large Europid data sets. They also showed that this risk ratio estimate was fully consistent with all previously published data on this variant, even though most of those previous – smaller – studies had reported no association (presumably due to type 2 error).

These, and other studies, have led to a concerted re-evaluation of the principles of candidate gene association studies in multifactorial disease and the promulgation of improved, more exacting 'industry standards' designed to deliver more robust results (Editorial 1999, Altshuler *et al.* 2000*a*, Cardon & Bell 2001). These include the need for: (i) significantly increased sample sizes (thousands of subjects, even more if gene–gene and gene–environment interactions are to be detected); (ii) incorporation of diverse study designs including case–control, family-based association studies and intermediate phenotype data sets, given the particular strengths and weaknesses of each approach (Cardon & Bell 2001); (iii) replication of findings in additional study groups of similar ethnic origin, and the exploitation of data sets from disparate ethnicities to unravel complex LD relationships between variants (Horikawa *et al.* 2000); (iv) an increasing emphasis on 'gene-wide' analyses including a full inventory of perigenic variation and comprehensive evaluation of association with disease, especially if the aim is definitive 'exclusion' of a gene from disease involvement; and (v) functional assessment of presumed aetiological variants, e.g. through *in vitro* or transgenic assays, to provide biological substantiation of statistical findings.

To date, very few studies of candidate genes for complex trait loci come close to approaching these requirements, leaving a slew of previous reports of association 'in limbo'. Examples from the T2D literature include associations with the genes for insulin (Bennett & Todd 1996, Huxtable *et al.* 2000), the sulphonylurea receptor (Inoue *et al.* 1996, Hani *et al.* 1998, 't Hart *et al.* 1999) and insulin receptor substrate 1 (Almind *et al.* 1993, Clausen *et al.* 1995, Hitman *et al.* 1995). A re-evaluation of some of these 'classic' candidates is therefore opportune. Crucially, the same experimental standards must be observed for all candidate genes, however they come to attention, including those defined initially on positional grounds (see below), and those arising out of more sophisticated and comprehensive assessments of biological candidacy, through expression profiling and proteomics, for example.

Analyses of animal and human models of disease

Given the intrinsic difficulties associated with a direct assault on the complex multifactorial traits themselves, one attractive strategy is to focus on more genetically tractable 'models' of those diseases, on the basis that genes identified in these models will provide clues to pathways implicated in the commoner, multifactorial forms of human disease. There are the following three main study options.

Study of monogenic forms of disease

Apposite examples include the analysis of maturity-onset diabetes of the young (MODY), an autosomal dominant, early-onset form of T2D (Hattersley *et al.* 1992) and the identification of single-gene effects underlying early-onset obesity (Montague *et al.* 1997). In the case of MODY, a combination of classical Mendelian linkage-based positional cloning approaches, and candidate gene studies, have revealed at least five different genes responsible for severe pancreatic beta-cell dysfunction and consequent diabetes (the genes for glucokinase (Froguel *et al.* 1992, Hattersley *et al.* 1992), hepatocyte-nuclear factors 1α (Yamagata *et al.* 1996*a*), 4α (Yamagata *et al.* 1996*b*) and 1β (Horikawa *et al.* 1997), and insulin promoter factor-1 (*IPF1*) (Stoffers *et al.* 1997)). These studies have provided valuable insights into the molecular circuitry of the beta-cell. With the exception of

IPF1, where mutations in the coding regions can, depending on the severity of functional impairment, result in either MODY or an increased predisposition to multifactorial T2D (Stoffers *et al.* 1998, Hani *et al.* 1999, MacFarlane *et al.* 1999), these particular genes do not seem to play a significant role in the later-onset forms of T2D. In the second example, identification of families segregating severe early-onset obesity due to mutations in the genes for leptin (Montague *et al.* 1997), the leptin receptor (Clément *et al.* 1998), the melanocortin-4 receptor (Vaisse *et al.* 1998, Yeo *et al.* 1998, Farooqi *et al.* 2000) and pro-opiomelanocortin (Krude *et al.* 1998) have confirmed the physiological role of these molecules in the control of energy balance in man.

Study of syndromic forms of disease

Common traits are sometimes observed as components within larger monogenic disease syndromes, e.g. T2D in partial lipodystrophy (Shackleton *et al.* 2000) or Friedreich's ataxia (Ristow *et al.* 1998); obesity in Bardet–Biedl syndrome (Katsanis *et al.* 2000). Gene identification for the syndrome (generally amenable to standard 'Mendelian' positional cloning methods) is clearly relevant to efforts to understand the pathogenesis of the associated complex trait.

Study of animal models of disease

Genetic dissection of relevant animal models is facilitated by a variety of factors, including large litter size, short generation times, capacity to engineer crosses and generate congenic lines and the ability to control environmental co-factors. Selective breeding, gene-targeting strategies and mutagenesis programmes have made available a wide range of rodent models for many diseases (Brown & Nolan 1998). In the T2D field, genetic dissection of polygenic models such as the Goto–Kakizaki (GK) rat is likely to be particularly relevant to human disease (Galli *et al.* 1996, Gauguier *et al.* 1996). At least seven loci controlling T2D-related subphenotypes have been identified in this model, revealing complex relationships between genotype and phenotype (for example, different loci influence fasting and post-load glycaemia) (Galli *et al.* 1996, Gauguier *et al.* 1996). At least one of these loci (*Nidd/gk2*) corresponds to a region implicated in human T2D susceptibility (chromosome 1q24) (Hanson *et al.* 1998, Elbein *et al.* 1999, Vionnet *et al.* 2000, Wiltshire *et al.* 2001).

Positional cloning in multifactorial disease

All the approaches described above rely on implicit assumptions. For candidate genes studies, the assumption is that the major genes influencing susceptibility to the disease of interest act in known biological pathways. For model-based approaches, the expectation is that findings from human and/or animal models of disease will be relevant to the multifactorial forms of disease. Although both assertions are perfectly reasonable, positional cloning methods, applied directly to families segregating multifactorial forms of disease, provide a means to progress gene discovery that is not hamstrung by such prior assumptions.

The objectives here are to apply linkage to define, then LD-based approaches to refine, disease-gene location. Although, in principle, it is a strength of such analyses that they can proceed without reference to the biology of the disease itself or to gene function, in practice, an assessment of the biological candidacy of the genes mapping into a region of interest defined by linkage is an important component of the strategy (the label 'positional candidate' analysis neatly describes this integration of positional and biological information (Collins 1995)).

Lessons from T1D

Theoretical difficulties inherent in undertaking such studies in multifactorial traits have been outlined above and are amply reinforced by data emerging from studies of the two main forms of diabetes. Although both T1D and T2D show strong familial aggregation, the extent of the familial clustering differs. The sibling relative risk (λ_s: the ratio of the risk of disease in the sibling of an affected individual compared with the population risk (Risch 1990)) for T1D in Europeans populations is of the order of 15 (6% risk in siblings: 0·4% in the population), but only $\sim$3·5 for T2D (35 vs 10%) (Köbberling & Tillil 1982, Vyse & Todd 1996). Since this index of familiarity sets an upper limit on the combined effect of all susceptibility genes (plus any component of

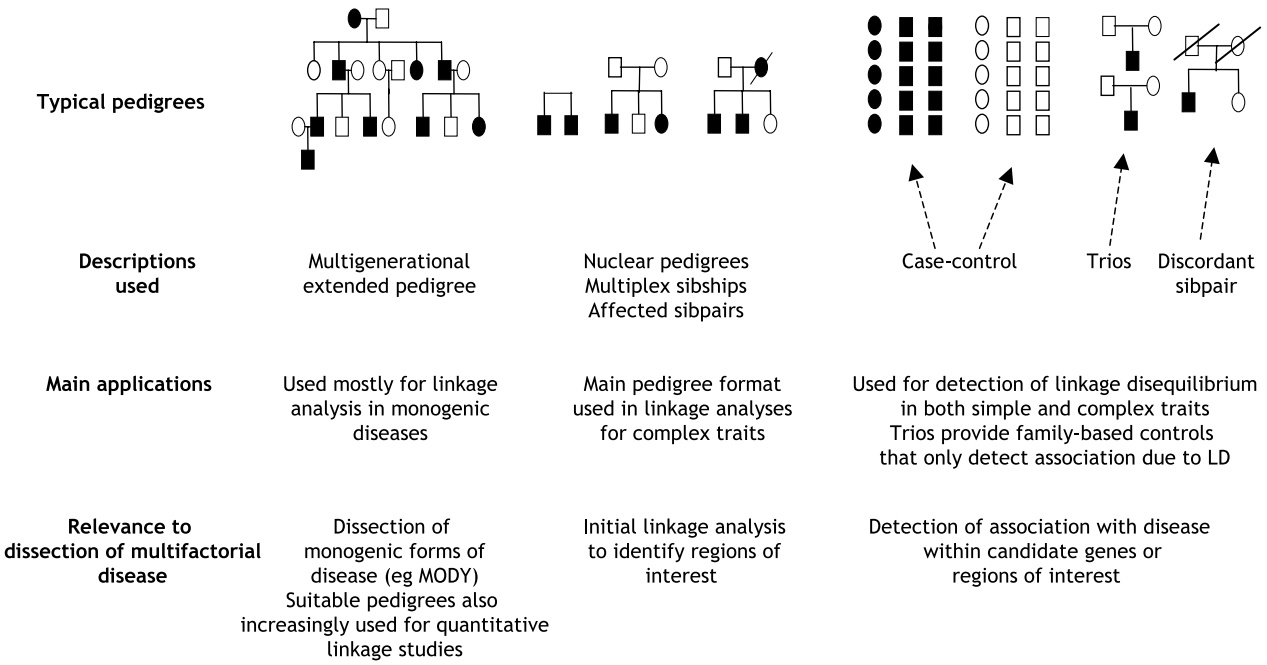

Figure 2 Configurations of some of the pedigree types used in dissection of multifactorial traits.

shared family environment), the power of linkage approaches was always likely to be greater for T1D than T2D (for equivalent sample size).

Indeed, the first successful applications of genome-wide linkage analysis to a complex multifactorial trait were described in T1D (Davies *et al.* 1994, Hashimoto *et al.* 1994). Because of the complex segregation patterns typical of multifactorial traits, the approach taken (and closely followed for many other complex traits) was based around the analysis of large numbers of affected sibling pairs (Fig. 2). Non-parametric (model-independent) linkage methods were used (Kruglyak *et al.* 1996) to identify those chromosomal regions where siblings sharing disease showed greater genetic similarity than expected by chance (Fig. 3).

These studies identified one very clear signal, around the HLA region on chromosome 6, accounting for about 40% of the inherited component of disease susceptibility. This result was not a surprise given the strong existing evidence for association between HLA alleles and T1D, but did provide a powerful validation of the methodology. The initial study (Davies *et al.* 1994) also threw up a number (around ten) of lesser signals, several of which have now been confirmed in other genome scans in T1D families (Concannon *et al.* 1998, Mein *et al.* 1998). Two of the regions contain strong candidate genes – those for *INS*, the gene for insulin, on chromosome 11p (Bennett & Todd 1996) and *CTLA4* on chromosome 2q (Marron *et al.* 1997) – which association studies have shown are both clearly implicated in disease susceptibility. Efforts continue to identify susceptibility genes underlying some of the other linkage signals, and to fill in the remaining pieces of the molecular puzzle. The picture emerging from these genetic studies neatly coincides with our evolving understanding of T1D pathogenesis, identifying variation in genes responsible for the regulation of the immune response to beta-cell antigens (*HLA* and *CTLA4*) or the level of thymic expression of those antigens (*INS*) as key determinants of inherited susceptibility to disease (Todd 1999).

T2D: a tougher target

The relatively low λ_s for T2D was always going to make it a tougher 'nut' to crack, and not surprisingly, success in gene identification has been slower to arrive. One reason has been the inevitable delay associated with ascertainment of the many hundreds of families required to compensate for the modest relative risks expected

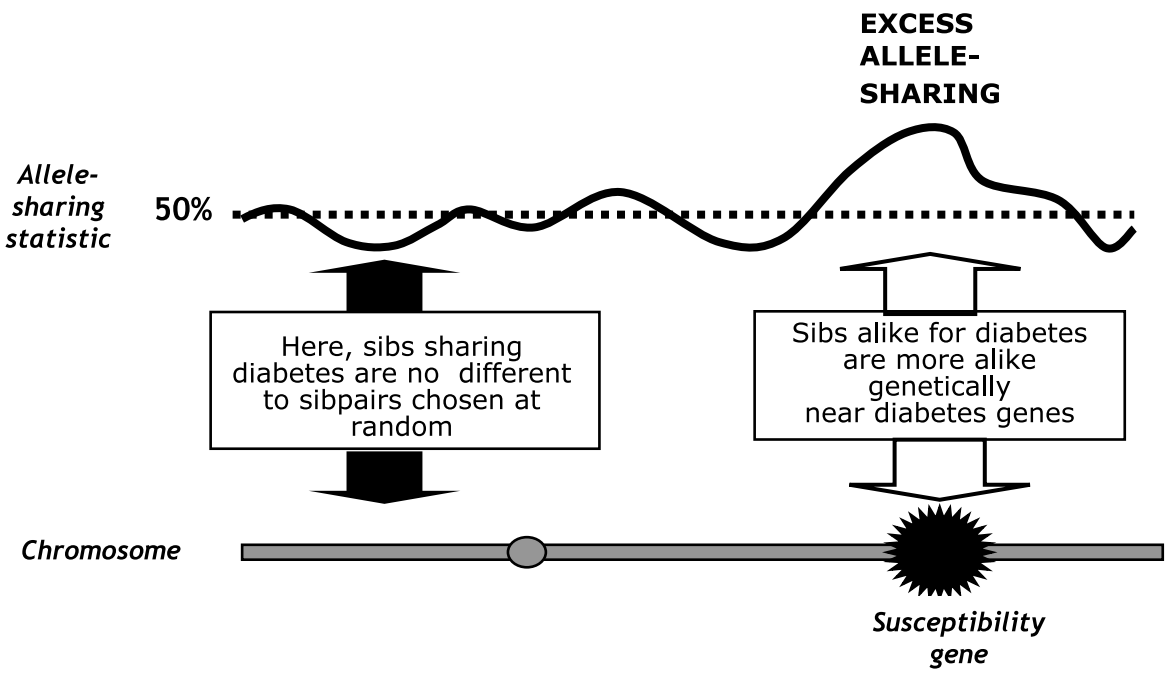

Figure 3 Principle behind non-parametric linkage analysis for a complex trait. The chromosome illustrated harbours a disease susceptibility gene at the position marked. A chromosome-wide scan has been performed on a set of T2D sibpairs. The result of that scan is summarised by an allele-sharing statistic (providing a measure of the percentage allele-sharing (identity by descent) at each point along the chromosome) which is shown above. At some distance from the susceptibility gene, siblings sharing disease are no more similar in terms of genotype than chance expectation and the allele-sharing statistic is close to 50% (stochastic variation means that it will vary around this level). However, near the gene of interest, siblings correlated for disease also show a correlation of their genotypes, reflecting the fact that they are more likely to have co-inherited a susceptibility allele from one (or both) parents. The peak of excess allele-sharing provides the clue that a susceptibility gene lies in the vicinity.

of any individual T2D susceptibility locus. Several large genome-wide scans for linkage have been completed for T2D in recent years, in a wide variety of populations, the largest in Pima Indians (Hanson *et al.* 1998), Finns (Mahtani *et al.* 1996, Ghosh *et al.* 2000, Watanabe *et al.* 2000), French (Vionnet *et al.* 2000), British (Wiltshire *et al.* 2001), Ashkenazim (Permutt *et al.* 2001) and Mexican-American (Hanis *et al.* 1996, Duggirala *et al.* 1999) pedigrees. The one absolutely clear conclusion from these results is that there is no single locus for T2D of major global significance equivalent to the contribution made by HLA to T1D susceptibility. However, as more data from these scans become public, it is reassuring to observe certain chromosomal regions emerging repeatedly; the most promising replicated signals are those on chromosomes 1, 12 and 20 (Ehm *et al.* 2000). For example, a 30 cM region centred on chromosome 1q24 has shown evidence for linkage to T2D in published data from Pima Indians (Hanson *et al.* 1998), French (Vionnet *et al.* 2000) and Utah Mormon (Elbein *et al.* 1999) studies, and is also replicated in the large UK 'Warren 2' study (573 affected sibpair families) (Wiltshire *et al.* 2001). As described above, further support for this locus is derived from the mapping of a diabetes-susceptibility locus to the equivalent region in the GK rat (Galli *et al.* 1996, Gauguier *et al.* 1996). This example also illustrates the important point that the evidence supporting the candidacy of a region often comes from various different sources and that, whilst guidelines for interpretation of linkage studies are essential (Lander & Kruglyak 1995), the case for a given region cannot always be distilled into a single significance value.

Clearly, in the face of the relatively modest signals for linkage expected in the analysis of multifactorial traits, replication of this kind can provide an important means of distinguishing those

peaks which are 'real' (i.e. harbouring susceptibility loci) from those likely to be 'spurious' (reflecting stochastic variation in the linkage statistic) (Lander & Kruglyak 1995). Nevertheless, it is important to appreciate the limitations of replication. There are several valid reasons why a real locus may prove hard to replicate, including ethnic heterogeneity (a locus may be more important in one particular population (Horikawa *et al.* 2000)), differences in diagnostic criteria or ascertainment scheme (McCarthy *et al.* 1998), and the effects of random variation on the power to detect a locus (Suarez *et al.* 1994).

There are several useful statistical tools which, by allowing more comprehensive examination of available data, should assist in this vital discrimination between real and spurious signals. These include stratification analyses, which, by reanalysing genotype data after stratification for relevant intermediate traits (e.g. age of disease onset, obesity), aim to minimise any loss of power associated with genetic heterogeneity (Merette *et al.* 1992, Watanabe *et al.* 1999, 2000). Conditional analyses, which set out to allow explicitly for the oligogenic aetiology of complex traits by seeking statistically significant (and therefore, by inference, biologically significant) interactions between regions of interest identified on genome scans, can also provide support for the biological relevance of regions showing evidence for linkage (Cox *et al.* 1999, Leal & Ott 2000). Finally, the development of much-improved computational and statistical methods for the analysis of continuous phenotypes in large pedigrees has facilitated a direct assault on the genetic dissection of those intermediate traits (for example, insulin sensitivity and beta-cell function in the case of T2D) considered to underlie the development of the dichotomous disease phenotype (Almasy & Blangero 1998, Duggirala *et al.* 1999). Through application of these methods, it should be possible to increase confidence that a given region emerging from a genome scan truly harbours a susceptibility gene, and that further positional cloning efforts are justified.

The post-genomic scan challenge

This 'locus validation' process represents only the initial step in gene discovery. The regions arising from a typical multifactorial trait genome scan are large (10–30 cM) and the peak of linkage in most cases provides only an approximate indication of the position of the susceptibility gene (Kruglyak & Lander 1995, Roberts *et al.* 1999). Whilst in Mendelian disease it is generally possible to narrow the critical region by typing additional meioses (if available), this is a highly inefficient procedure in complex traits, where the poor correlation between genotype and phenotype means that any individual recombination event carries only limited (statistical) information on the location of the disease gene.

The researcher aiming to positionally clone a complex trait gene is therefore faced typically with the daunting task of addressing a region of approximately 20 cM, likely to contain ~20 million bases, and 200 or more genes. This region will contain about 60 000 common variants (mostly single-nucleotide polymorphisms (SNPs)) of which up to 1000 will be in coding sequence, and as many as another 15 000–20 000 in sequences potentially relevant to gene function and regulation (introns, untranslated regions, promoters, remote regulatory regions). The task of identifying the single variant (or set of variants) that confers susceptibility remains a major endeavour.

Essentially, there are two complementary, interrelated approaches that are currently applicable. The prospects for both have been very significantly enhanced by access to the increasingly complete human genome sequence (International Human Genome Sequencing Consortium 2001) and related efforts to annotate that sequence to identify the location of expressed sequences (Birney *et al.* 2001, Shoemaker *et al.* 2001) and common sites of human variation (The International SNP Map Working Group 2001).

The first approach relies on LD mapping to improve localisation within the region of linkage. As described earlier, the fact that LD extends only a relatively short distance (tens of kilobases is a reasonable estimate for outbred populations) from the susceptibility locus should mean that LD is capable of reducing substantially the interval of interest (Lonjou *et al.* 1999, Abecasis *et al.* 2001, Reich *et al.* 2001). Having said that, the highly unpredictable pattern and extent of LD in any given situation means that any systematic search across a large genomic region is likely to require a very high density of markers (arguably, at least one every 10 kb) and, because the effect sizes to be detected are modest, rather large sample sizes (hundreds, even thousands, of subjects) (Kruglyak

1999, Roses 2000, Weiss & Terwilliger 2000). Whilst the markers are now available, with several million SNPs catalogued in public and proprietary databases (The International SNP Map Working Group 2001), these prestigious genotyping requirements remain prohibitive with current technology on both economic and logistical grounds. Three developments are likely to ease the situation in the medium term. The first is the development of more robust, less-expensive, high-throughput methods for SNP typing (Kwok 2000). The second is the development and validation of methods for deriving reliable estimates of allele frequencies from pooled DNA samples, which, if successful, will reduce substantially the amount of genotyping involved in surveying a region for LD (Ross *et al.* 2000, Germer *et al.* 2000). The third, and a little way further in the future, is the elucidation and dissemination of genome-wide 'haplotype maps', providing access to the 'baseline terrain' of LD in any given region (and population) and thereby enhancing efforts to pick up departures from that baseline in samples of disease chromosomes (Service *et al.* 2001).

The complement to such 'indirect' LD mapping approaches to gene identification concentrates more on the detailed evaluation of the strongest positional candidates in the region of interest. The first step in such an analysis involves retrieval of as complete as possible a list of the genes within the region. Fortunately, the improving annotation of the draft human sequence is capable of delivering increasingly complete transcript inventories (Birney *et al.* 2001). Even so, the number of genes in a typical region of linkage (hundreds) is likely to be too large to contemplate analysing all of them for variation and association with disease, making some sort of prioritisation desirable to arrive at a 'shortlist' of the most promising positional candidates. Such a shortlist needs to match the known and presumed function of the various regional transcripts and their patterns of tissue expression to the researcher's knowledge of the disease of interest. The data informing this process may be derived from the burgeoning genomics databases (e.g. ENSEMBL, www.ensembl.org) and/or from in-house laboratory analyses (for example, determining the qualitative and quantitative expression profiles of regional candidates by interrogating a regional cDNA microarray with message from tissues of interest (Aitman *et al.* 1999, Ugolini *et al.* 1999, Shoemaker *et al.* 2001)).

Characterising candidate genes

Once a strong positional candidate has been identified, the final common pathway is well-travelled and essentially the same as the analysis of any candidate gene (see above). Relevant parts of the gene need to be resequenced to compile an inventory of genomic variation, and the variants uncovered tested for association with disease. Ideally, such studies should employ a combination of case–control and family-based association tests (Editorial 1999, Cardon & Bell 2001).

Several important considerations need to be emphasised. First, it seems likely that variants involved in complex trait susceptibility will often be acting through effects on transcriptional regulation and/or RNA stability (rather than through altered primary and secondary amino acid structure). Any comprehensive gene survey therefore needs to include 'unfashionable real estate' including untranslated regions, all intronic sequence and (given poor characterisation thus far of critical regulatory regions) a considerable stretch of upstream sequence. Secondly, the vagaries of LD mean that involvement of a gene in disease susceptibility cannot be excluded simply because a subset of variants in that gene show no association with disease; systematic and comprehensive analyses of variation in multiple populations are required. Thirdly, susceptibility may often be governed by the combined action of several different variants within a gene (each, for example, having a cumulative effect on transcriptional activity or RNA stability). Several likely examples of this exist including the calpain-10 gene (*CAPN10*) in T2D (Horikawa *et al.* 2000); and insulin (*INS*) (Bennett & Todd 1996, Stead *et al.* 2000) and *HLA* in T1D (Zavattari *et al.* 2001). Such complex intra-locus interactions certainly complicate interpretation of association data. Finally, the genetic architecture of complex traits remains uncertain. If the 'common disease, common variant' hypothesis holds up (Cargill *et al.* 1999, The International SNP Map Working Group 2001), this should mean that the extent of allelic heterogeneity is relatively modest, and facilitates both the LD mapping and the consequent functional characterisation of candidate variants. If, instead, allelic heterogeneity is more

widespread, and susceptibility to a given trait more often determined by multiple, diverse, low-frequency susceptibility alleles (Pritchard 2001), LD will be harder to find, and confirming the functional relevance of any single variant more problematical (Todd 2001).

Recent successes in positional cloning

It might be tempting to conclude from the above that the prospects for successful gene identification by positional cloning in complex traits are poor. Recent successes in both T2D (Horikawa *et al.* 2000, Roses *et al.* 2000) and Crohn's disease (Hugot *et al.* 2001, Ogura *et al.* 2001) strongly suggest otherwise, and demonstrate the ways in which the approaches described above have been successfully applied to gene discovery.

Hanis *et al.* (1996) reported their genome scan on 258 Mexican-American sibships, which provided significant evidence of linkage to T2D around the marker D2S125 on chromosome 2q (designated *NIDDM1*). Attempts to replicate this finding in other ethnic groups were mostly unsuccessful (Hani *et al.* 1997, Ghosh *et al.* 1998). However, using the conditional analysis methods alluded to above, independent support for *NIDDM1* was obtained by demonstrating significant interaction between *NIDDM1* and a second region on chromosome 15 which had produced a modest signal in the original genome scan (Cox *et al.* 1999). This observation clearly required both regions to be having some biological effect on disease susceptibility. These conditional analyses also helped to refine the 'confidence interval' of *NIDDM1* to around 2 Mb (equivalent to 7 cM in this telomeric location). At this point, the team switched to an LD-based approach (Horikawa *et al.* 2000), identifying SNPs in the region and testing them for association with T2D. Some initial hints of LD focused their attention on a 66 kb region containing three genes, and this was targeted for exhaustive variant detection and further association analyses. A variant in intron 3 of the calpain-10 gene (UCSNP-43) emerged with the best statistical credentials from these analyses; it had the strongest association, and it successfully partitioned the evidence for linkage in the original scan. There were, however, three concerns about this SNP (Altshuler *et al.* 2000*b*). First, it was intronic and did not appear to influence splicing. Secondly, homozygotes for the at-risk allele at UCSNP-43 were, rather surprisingly, not at increased risk. Thirdly, the at-risk allele was highly prevalent (75%) and could not, in isolation, explain the size of the linkage signal. Could it be that other SNPs were contributing to the susceptibility effect? Haplotype studies within *CAPN10* in Mexican-American and other European populations (Horikawa *et al.* 2000) suggest that this is indeed the case, and that individual risk is best described by the individual configuration of alleles at a number of variant sites within the gene.

Final confirmation that these *CAPN10* variants are functional will require biological rather than statistical enquiry, for example, through examination of calpain-10 (tissue-specific) knockout mice. Such biological verification should expunge any residual claims that the true aetiological variant lies, undiscovered, elsewhere in the region, its signal being detectable at *CAPN10* due to extensive LD relationships across chromosome 2q (Altshuler *et al.* 2000*b*). These functional studies should also start to address the subsequent question: how is that variation in this ubiquitously expressed protease influences risk of diabetes?

Another triumph of the approaches described, outside the metabolic arena, has been the identification of variants in the *NOD2* gene as the basis for the region of linkage to Crohn's disease previously identified on chromosome 16 (*IBD1*). Interestingly, the two groups that reported this finding (Hugot *et al.* 2001, Ogura *et al.* 2001) used rather different approaches to arrive at this discovery, corresponding in fact, to the two main strategies described earlier (Todd 2001). One group (Hugot *et al.* 2001) adopted a strategy based on indirect LD mapping. This allowed them (rather serendipitously, it must be said) to localise the susceptibility gene to a much smaller region and directed their focus towards the small subset of genes it contained, *NOD2* amongst them. The other successful group (Ogura *et al.* 2001) embarked on a positional candidate approach and moved directly to a detailed analysis of *NOD2* as soon as data emerged supporting its biological candidacy in the disease of interest.

What does the future hold?

There is no doubt, looking back on the last decade, that most efforts to map complex trait susceptibility

genes have been fuelled by a heady mix of optimism and a rather unsophisticated attitude to the complexities of multifactorial disease. Quite simply, tools which were proving increasingly successful at delivering gene identification in monogenic diseases were not up to the task of scaling the more demanding heights of complex trait gene discovery. This deficit is now clearly being remedied by a host of technical and biological advances, availability of (as yet, incomplete) annotated human sequence being perhaps the most spectacular example. The recent clatter of complex trait genes identified using the approaches described promises to herald a new wave of success in understanding these major diseases.

What further advances can we expect to expedite and support these future successes? First, improved methods for defining biological candidacy and in the analysis of complex biological systems will undoubtedly speed positional candidate selection and advance efforts to characterise disease pathogenesis once susceptibility genes are found. These analyses will depend on the capacity to conduct global ('genome-wide') analyses at various levels of cellular and organismal organisation – including the transcriptome, proteome and metabonome – and the ability to integrate these disparate information sources (Vidal 2001). Secondly, the field needs more powerful (technical and statistical) tools for LD mapping in large populations. These should allow researchers to extract more of the information contained in population resources. Given the relatively low power of linkage approaches within families, there is no doubt that LD analyses will be required to take over where linkage alone is likely to fail, in localising signals within large genomic regions, and/or detecting genes of lesser effect (Risch & Merikangas 1996). Ultimately, these advances may permit genome-wide scans for LD to become practicable. At the same time, we need improved statistical tools for determining which variants within a tract of associated polymorphisms are most likely to be functional. Finally, we need access to very large, well-characterised populations so that gene–gene and gene–environment interactions can be explored with the necessary rigour and power. This will provide the most complete, integrated view of the factors determining disease risk and pathogenesis, and their interactions, and lay the platform for the robust specification of individual risk profile and prognosis, the latter clearly a pre-requisite for future efforts to develop personalised health care

With these and other advances we can expect the next decade to see many more complex traits yield their secrets to the gene mappers.

References

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattachararyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF & Cookson WO 2001 Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics* **68** 191–197.

Aitman TJ, Glazier AM, Wallace CA, Cooper LD, Norsworthy PJ, Wahid FN, Al-Majali KM, Trembling PM, Mann CJ, Shoulders CC, Graf D, St Lezin E, Kurtz TW, Kren V, Pravenec M, Ibrahimi A, Abumrad NA, Stanton LW & Scott J 1999 Identification of Cd36 Fat as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. *Nature Genetics* **21** 76–83.

Almasy L & Blangero J 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62** 1198–1211.

Almind K, Bjørbaek C, Vestergaard H, Hansen T, Echwald S & Pedersen O 1993 Amino acid polymorphisms of insulin receptor substrate-1 in non-insulin dependent diabetes mellitus. *Lancet* **342** 828–832.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L & Lander ES 2000*a* The common PPARgamma Pro12 Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* **26** 76–80.

Altshuler D, Daly M & Kruglyak L 2000*b* Guilt by association. *Nature Genetics* **26** 135–137.

Bennett ST & Todd JA 1996 Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Annual Review of Genetics* **30** 343–370.

Birney E, Bateman A, Clamp ME & Hubbard TJ 2001 Mining the human draft genome. *Nature* **409** 827–828.

Boehnke M & Langefeld CD 1998 Genetic association mapping based on discordant sib pairs: the discordant alleles test. *American Journal of Human Genetics* **62** 950–961.

Brown SDM & Nolan PM 1998 Mouse mutagenesis – systematic studies of mammalian gene function. *Human Molecular Genetics* **7** 1627–1633.

Cardon LR & Bell JI 2001 Association study designs for complex disease. *Nature Reviews in Genetics* **2** 91–99.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ & Lander ES 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22** 231–238.

Clausen JO, Hansen T, Bjorbaek C, Echwald SM, Urhammer SA, Rasmussen S, Andersen CB, Hansen L, Almind K, Winther K, Haraldsdòttir J, Borch-Johnsen K & Pedersen O 1995 Insulin resistance: interactions between obesity and a common variant of insulin receptor substrate-1. *Lancet* **346** 397–402.

Clément K, Vaisse C, Lahlou N, Cabrol S, Pelloux V, Cassuto D, Gourmelen M, Dina C, Chambaz J, Lacorte JM, Basdevant A, Bougnè res P, Lebouc Y, Froguel P & Guy-Grand B 1998 A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392** 398–401.

Collins FS 1995 Positional cloning moves from the perditional to traditional. *Nature Genetics* **9** 347–350.

Concannon P, Gogolin-Ewens KJ, Hinds DA, Wapelhorst B, Morrison VA, Stirling B, Mitra M, Farmer J, Williams SR, Cox NJ, Bell GI, Risch N & Spielman RS 1998 A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nature Genetics* **19** 292–296.

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI & Kong A 1999 Loci on chromosomes 2 NIDDM1 and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics* **21** 213–215.

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC & Todd JA 1994 A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* **371** 130–136.

Duggirala R, Blangero J, Almasy L, Dyer TD, Williams KL, Leach RJ, O'Connell P & Stern MP 1999 Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *American Journal of Human Genetics* **64** 1127–1140.

Editorial 1999 Freely associating. *Nature Genetics* **22** 1–2.

Ehm MG, Kkarnoub MC, Sakul H, Gottschalk K, Holt DC, Weber JL, Vaske D, Briley D, Briley L, Kopf J, McMillen P, Nguyen Q, Reisman M, Lai EH, Joslyn G, Shepherd NS, Bell C, Wagner MJ, Burns DK & The American Diabetes Association GENNID Study Group 2000 Genome-wide search for type 2 diabetes susceptibility genes in four American populations. *American Journal of Human Genetics* **66** 1871–1881.

Elbein SC, Hoffman MD, Teng K, Leppert MF & Hasstedt SJ 1999 A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. *Diabetes* **48** 1175–1182.

Farooqi IS, Yeo GS, Keogh JM, Aminian S, Jebb SA, Butler G, Cheetham T & O'Rahilly S 2000 Dominant and recessive inheritance of human obesity associated with melanocortin 4 receptor deficiency. *Journal of Clinical Investigation* **106** 271–279.

Frayling TM, Walker M, McCarthy MI, Evans JC, Ayres S, Allen LI, Ellard S, Lynn S, Turner RC, O'Rahilly S, Hitman GA & Hattersley AT 1999 Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* **48** 2475–2479.

Froguel P, Vaxillaire M, Sun F, Velho G, Zouali H, Butel MO, Lesage S, Vionnet N, Clément K, Fougerousse F, Tanizawa Y, Weissenbach J, Beckmann JS, Lathrop GM, Passa Ph, Permutt MA & Cohen D 1992 Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature* **356** 162–165.

Galli J, Li LS, Glaser A, Östenson CG, Jiao H, Fakhrai-Rad H, Jacob HJ, Lander ES & Luthman H 1996 Genetic analysis of non-insulin dependent diabetes mellitus in the GK rat. *Nature Genetics* **12** 31–37.

Gauguier D, Froguel P, Parent V, Bernard C, Bihoreau MT, Portha B, James MR, Penicaud L, Lathrop M & Ktorza A 1996 Chromosomal mapping of genetic loci associated with non-insulin dependent diabetes in the GK rat. *Nature Genetics* **12** 38–43.

Germer S, Holland MJ & Higuchi R 2000 High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Research* **10** 258–266.

Ghosh S & Schork NJ 1996 Genetic analysis of NIDDM. The Study of quantitative traits. *Diabetes* **45** 1–14.

Ghosh S, Hauser ER, Magnuson VL, Valle T, Ally DS, Karanjawala ZE, Rayman JB, Knapp JI, Musick A, Tannenbaum J, Te C, Eldridge W, Shapiro S, Musick T, Martin C, So A, Witt A, Harvan JB, Watanabe RM, Hagopian W, Eriksson J, Nylund SJ, Kohtamaki K, Tuomilehto-Wolf E, Toivanen L, Vidgren G, Ehnholm C, Bergman RN, Tuomilehto J, Collins FS & Boehnke M 1998 A large sample of Finnish diabetic sib-pairs reveals no evidence for a non-insulin-dependent diabetes mellitus susceptibility locus at 2 qter. *Journal of Clinical Investigation* **102** 704–709.

Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, Mohlke KL, Silander K, Kohtamaki K, Chines P, Balow J Jr, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kudelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB, Rha D, Segal L, Shapiro S, Sharaf R, Shurtleff B, So A, Tannenbaum J, Te C, Tovar J, Unni A, Welch C, Whiten R, Witt A, Blaschak-Harvan J, Douglas JA, Duren WL, Epstein MP, Fingerlin TE, Kaleta HS, Langer EM, Li C, McEachin RC, Stringham HM, Trager E, White PP, Eriksson J, Toivanen L, Vidgren G, Nylund SJ, Tuomilehto-Wolf E, Ross EH, Demirchyan E, Hagopian WA, Buchanan TA, Tuomilehto J, Bergman RN, Collins FS & Boehnke M 2000 The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics FUSION study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *American Journal of Human Genetics* **67** 1174–1185.

Hani EH, Hager J, Philippi A, Demenais F, Froguel P & Vionnet N 1997 Mapping NIDDM susceptibility loci in French families: studies with markers in the region of NIDDM1 on chromosome 2q. *Diabetes* **46** 1225–1226.

Hani EH, Boutin P, Durand E, Inoue H, Permutt MA, Velho G & Froguel P 1998 Missense mutations in the pancreatic islet beta cell inwardly rectifying K^+ channel gene *KIR6·2/BIR* : a meta-analysis suggests a role in the polygenic basis of type II diabetes mellitus in Caucasians. *Diabetologia* **41** 1511–1515.

Hani EH, Stoffers DA, Chè vre J-C, Durand E, Stanojevic V, Dina C, Habener JF & Froguel P 1999 Defective mutations in the insulin promoter factor-1 *IPF-1* gene in late-onset type 2 diabetes mellitus. *Journal of Clinical Investigation* **104** R41–R48.

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, Wapelhorst B, Spielman RS, Gogolin-Ewens KJ, Shephard JM, Williams SR, Risch N, Hinds D, Iwasaki N, Ogata M, Omori Y, Petzold C, Rietzsch H, Schroder HE, Schulze J, Cox NJ, Menzel S, Boriraj VV, Chen X, Lim LR, Lindner T, Mereu LE, Wang YQ, Xiang K, Yamagata K, Yang Y & Bell GI 1996 A genome-wide search for human non-insulin-dependent type 2 diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genetics* **13** 161–171.

Hanson RL, Ehm MG, Pettitt DJ, Prochazka M, Thompson DB, Timberlake D, Foroud T, Kobes S, Baier L, Burns DK, Almasy L, Blangero J, Garvey WT, Bennett PH & Knowler WC 1998 An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *American Journal of Human Genetics* **63** 1124–1132.

't Hart LM, De Knijff P, Dekker JM, Stolk RP, Nijpels G, Van der Does FE, Ruige JB, Grobbee DE, Heine RJ & Maassen JA 1999 Variants in the sulphonylurea receptor gene: association of the exon 16–3t variant with type II diabetes mellitus in Dutch Caucasians. *Diabetologia* **42** 617–620.

Hashimoto L, Habita C, Beressi JP, Delepine M, Besse C, Cambon-Thomsen A, Deschamps I, Rotter JP, Djoulah S, Froguel P, Weissenbach J, Lathrop GM & Julier C 1994 Genetic mapping of a multifactorial disease: evidence for a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q13. *Nature* **371** 161–164.

Hattersley AT, Turner RC, Permutt MA, Patel P, Tanizawa Y, Chiu KC, O'Rahilly S, Watkins PJ & Wainscoat JS 1992 Linkage of type 2 diabetes to the glucokinase gene. *Lancet* **339** 1307–1310.

Hitman GA, Hawrami K, McCarthy MI, Viswanathan M, Snehalatha C, Ramachandran A, Tuomilehto J, Tuomilehto-Wolf E, Nissinen A & Pedersen O 1995 Insulin receptor substrate-1

gene mutations in NIDDM; implications for the study of polygenic disease. *Diabetologia* **38** 481–486.

Horikawa Y, Iwasaki N, Hara M, Furuta H, Hinokio Y, Cockburn BN, Lindner T, Yamagata K, Ogata M, Tomonaga O, Kuroki H, Kasahara T, Iwamoto Y & Bell GI 1997 Mutation in hepatocyte nuclear factor-1β gene TCF2 associated with MODY. *Nature Genetics* **17** 384–385.

Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TM, Mashima H, Schwarz PEH, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL & Bell GI 2000 Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* **26** 163–175.

Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cézard J-P, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel J-F, Sahbatou M & Thomas G 2001 Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411** 599–603.

Huxtable SJ, Saker PJ, Haddad L, Walker M, Frayling TM, Levy JC, Hitman GA, O'Rahilly S, Hattersley AT & McCarthy MI 2000 Analysis of parent-offspring trios provides evidence for linkage and association between the insulin gene and type 2 diabetes mediated exclusively through paternally transmitted class III variable number tandem repeat alleles. *Diabetes* **49** 126–130.

Inoue H, Ferrer J, Welling CM, Elbein SC, Hoffman M, Mayorga R, Warren-Perry M, Zhang Y, Millns H, Turner R, Province M, Bryan J, Permutt MA & Aguilar-Bryan L 1996 Sequence variants in the sulfonylurea receptor SUR gene are associated with NIDDM in Caucasians. *Diabetes* **45** 825–831.

International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409** 860–921.

The International SNP Map Working Group 2001 A map of human genome sequence variation containing 1·42 million single nucleotide polymorphisms. *Nature* **409** 928–933.

Jorde LB 1995 Linkage disequilibrium as a gene-mapping tool. *American Journal of Human Genetics* **56** 11–14.

Jorde LB 2000 Linkage disequilibrium and the search for complex disease genes. *Genome Research* **10** 1435–1444.

Katsanis N, Beales PL, Woods MO, Lewis RA, Green JS, Parfrey PS, Ansley SJ, Davidson WS & Lupski JR 2000 Mutations in MKKS cause obesity, retinal dystrophy and renal malformations associated with Bardet–Biedl syndrome. *Nature Genetics* **26** 67–70.

Keavney B, McKenzie C, Parish S, Palmer A, Clark S, Youngman L, Delepine M, Lathrop M, Peto R, Collins R for the ISIS Collaborators 2000 Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. *Lancet* **355** 434–442.

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M & Tsui LC 1989 Identification of the cystic fibrosis gene: genetic analysis. *Science* **245** 1073–1080.

Kiesewetter S, Macek M Jr, Davis C, Curristin SM, Chu CS, Graham C, Shrimpton AE, Cashman SM, Tsui LC, Mickle J, Amos J, Highsmith WE, Shuber A, Witt DR, Crystal RG & Cutting GR 1993 A mutation in CFTR produces different phenotypes depending on chromosomal background. *Nature Genetics* **5** 274–278.

Köbberling J & Tillil H 1982 Empirical risk figures for first degree relatives of non-insulin-dependent diabetics. In *The Genetics of Diabetes Mellitus*, pp 201–209. Eds J Köbberling & R Tattersall. London: Academic Press.

Krude H , Biebermann H, Luck W, Horn R, Brabant G & Gruters A 1998 Severe early-onset obesity, adrenal insufficiency and red hair pigment caused by POMC mutations in humans. *Nature Genetics* **19** 155–157.

Kruglyak L 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22** 139–144.

Kruglyak L & Lander ES 1995 High-resolution genetic mapping of complex traits. *American Journal of Human Genetics* **56** 1212–1223.

Kruglyak L, Daly MJ, Reeve-Daly MP & Lander ES 1996 Parametric and non-parametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58** 1347–1363.

Kwok PY 2000 High-throughput genotyping assay approaches. *Pharmacogenomics* **1** 95–100.

Lander E & Kruglyak L 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11** 241–247.

Lander ES & Schork NJ 1994 Genetic dissection of complex traits. *Science* **265** 2037–2048.

Leal SM & Ott J 2000 Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. *American Journal of Human Genetics* **66** 567–575.

Lonjou C, Collins A & Morton NE 1999 Allelic association between marker loci. *PNAS* **96** 1621–1626.

McCarthy M & Hitman GA 1993 The genetics of non-insulin dependent diabetes mellitus. In *The Causes of Diabetes*, pp 157–186. Ed. RDG Leslie. London: John Wiley.

McCarthy MI, Kruglyak L & Lander ES 1998 Sib pair collection strategies for complex diseases. *Genetic Epidemiology* **15** 317–340.

MacFarlane WM, Frayling TM, Ellard S, Evans JC, Allen LIS, Bulman MP, Ayres S, Shepherd M, Clark P, Millward A, Demaine A, Wilkin T, Docherty K & Hattersley AT 1999 Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *Journal of Clinical Investigation* **104** R33–R39.

Mahtani MM, Widén E, Lehto M, Thomas J, McCarthy M, Brayer J, Bryant B, Chan G, Daly M, Forsblom C, Kanninen T, Kirby A, Kruglyak L, Munnelly K, Parkkonen M, Reeve-Daly MP, Weaver A, Brettin T, Duyk G, Lander ES & Groop LC 1996 Mapping of a gene for NIDDM associated with an insulin secretion defect by a genome scan in Finnish families. *Nature Genetics* **14** 90–95.

Marron MP, Raffel LJ, Garchon HJ, Jacob CO, Serrano-Rios M, Martinez Larrad MT, Teng WP, Park Y, Zhang ZX, Goldstein DR, Tao YW, Beaurain G, Bach JF, Huang HS, Luo DF, Zeidler A, Rotter JI, Yang MC, Modilevsky T, Maclaren NK & She JX 1997 Insulin-dependent diabetes mellitus (IDDM) is associated with CTLA4 polymorphisms in multiple ethnic groups. *Human Molecular Genetics* **6** 1275–1282.

Mein CA, Esposito L, Dunn MG, Johnson GC, Timms AE, Goy JV, Smith AN, Sebag-Montefiore L, Merriman ME, Wilson AJ, Pritchard LE, Cucca F, Barnett AH, Bain SC & Todd JA 1998 A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nature Genetics* **19** 297–300.

Merette C, King MC & Ott J 1992 Heterogeneity analysis of breast cancer families by using age at onset as covariate. *American Journal of Human Genetics* **50** 515–519.

Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, Wareham NJ, Sewter CP, Digby JE, Mohammed SN, Hurst JA, Cheetham CH, Earley AR, Barnett AH, Prins JB & O'Rahilly S 1997 Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387** 903–908.

Neel JV 1982 The thrifty genotype revisited. In *The Genetics of Diabetes Mellitus*, pp 283–293. Eds J Köbberling & R Tattersall. London: Academic Press.

Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar J-P,

Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nuñez G & Cho JH 2001 A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411** 603–607.

Ott J 1990 Cutting a Gordian knot in the linkage analysis of complex human traits. *American Journal of Human Genetics* **46** 219–221.

Ott J 1999 *Analysis of Human Genetic Linkage*. Baltimore: The Johns Hopkins University Press.

Peltonen L & McKusick V 2001 Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291** 1224–1229.

Permutt MA, Wasson JC, Suarez BK, Lin J, Thomas J, Meyer J, Lewitzky S, Rennich JS, Parker A, DuPrat L, Maruti S, Chayen S & Glaser B 2001 A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. *Diabetes* **50** 681–685.

Pritchard JK 2001 Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics* **69** 124–137.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R & Lander ES 2001 Linkage disequilibrium in the human genome. *Nature* **411** 199–204.

Risch N 1990 Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* **46** 222–228.

Risch N 2000 Searching for genetic determinants in the new millennium. *Nature* **405** 847–856.

Risch N & Merikangas K 1996 The future of genetic studies of complex human diseases. *Science* **273** 1516–1517.

Ristow M, Giannakidou E, Hebinck J, Busch K, Vorgerd M, Kotzka J, Knebel B, Mueller-Berghaus J, Epplen C, Pfeiffer A, Kahn CR, Doria A, Krone W & Mueller-Wieland D 1998 An association between NIDDM and a GAA trinucleotide repeat polymorphism in the X25/frataxin (Friedreich's ataxia) gene. *Diabetes* **47** 851–854.

Roberts SB, MacLean CJ, Neale MC, Eaves LJ & Kendler KS 1999 Replication of linkage studies of complex traits: an examination of variation in location estimates. *American Journal of Human Genetics* **65** 876–884.

Roses AD 2000 Pharmacogenetics and the practice of medicine. *Nature* **405** 857–865.

Ross P, Hall L & Haff LA 2000 Quantitative approach to single-nucleotide polymorphism analysis using MALDI-TOF mass spectrometry. *BioTechniques* **29** 620–626.

Service SK, Ophoff RA & Freimer NB 2001 The genome-wide distribution of background linkage disequilibrium in a population isolate. *Human Molecular Genetics* **10** 545–551.

Shackleton S, Lloyd DJ, Jackson SNJ, Evans R, Niermeijer MF, Singh BM, Schmidt H, Brabant G, Kumar S, Durrington PN, Gregory S, O'Rahilly S & Trembath RC 2000 *LMNA*, encoding lamin A/C, is mutated in partial lipodystrophy. *Nature Genetics* **24** 153–156.

Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G, Wu LF, Altschuler SJ, Edwards S, King J, Tsang JS, Schimmack G, Schelter JM, Koch J, Ziman M, Marton MJ, Li B, Cundiff P, Ward T, Castle J, Krolewski M, Meyer MR, Mao M, Burchard J, Kidd MJ, Dai H, Phillips JW, Linsley PS, Stoughton R, Scherer S & Boguski MS 2001 Experimental annotation of the human genome using microarray technology. *Nature* **409** 922–927.

Spielman RS & Ewens WJ 1996 The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59** 983–989.

Stead JDH, Buard J, Todd JA & Jeffreys AJ 2000 Influence of allele lineage on the role of the insulin minisatellite in susceptibility to type 1 diabetes. *Human Molecular Genetics* **9** 2929–2935.

Stoffers DA, Ferrer J, Clarke WL & Habener JF 1997 Early-onset type-II diabetes mellitus MODY4 linked to IPF1. *Nature Genetics* **17** 138–139.

Stoffers DA, Stanojevic V & Habener JF 1998 Insulin promoter factor-1 gene mutation linked to early-onset type 2 diabetes mellitus directs expression of a dominant negative isoprotein. *Journal of Clinical Investigation* **102** 232–241.

Suarez BK, Hampe CL & Van Eerdewegh P 1994 Problems of replicating linkage claims in psychiatry. In *Genetic Approaches to Mental Disorders*, pp 23–46. Eds ES Gershon & CR Cloninger. Washington DC: American Psychiatric Press.

Todd JA 1999 From genome to aetiology in a multifactorial disease, type 1 diabetes. *BioEssays* **21** 164–174.

Todd JA 2001 Tackling common disease. *Nature* **411** 537–539.

Tuomi T, Groop LC, Zimmet PZ, Rowley MJ, Knowles W & Mackay IR 1993 Antibodies to glutamic acid decarboxylase reveal latent autoimmune diabetes mellitus in adults with a non-insulin-dependent onset of disease. *Diabetes* **42** 359–362.

Ugolini F, Adélaïde J, Charafe-Jauffret E, Nguyen C, Jacquemier J, Jordan B, Birnbaum D & Pébusque M-J 1999 Differential expression array of chromosome arm 8p genes identifies Frizzled-related (FRP1/FRZB) and Fibroblast Growth Factor Receptor 1 (FGFR1) as candidate breast cancer genes. *Oncogene* **18** 1903–1910.

Vaisse C, Clément B, Grand-Guy B & Froguel P 1998 A frameshift mutation in human melanocortin-4 receptor results in a dominant form of obesity. *Nature Genetics* **20** 113–114.

Vidal M 2001 A biological atlas of functional maps. *Cell* **104** 333–339.

Vionnet N, Hani EH, Dupont S, Gallina S, Francke S, Dotte S, De Matos F, Durand E, Lepêtre F, Lecouer C, Gallina P, Zekiri L, Dina C & Froguel P 2000 Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21-q24. *American Journal of Human Genetics* **67** 1470–1480.

Vyse TJ & Todd JA 1996 Genetic analysis of autoimmune disease. *Cell* **85** 311–318.

Watanabe RM, Ghosh S, Birznieks G, Duren WL & Mitchell BD 1999 Application of an ordered subset analysis approach to the genetics of alcoholism. *Genetic Epidemiology* **17** (Suppl 1) S385–S390.

Watanabe RM, Ghosh S, Langefeld CD, Valle TT, Hauser ER, Magnuson VL, Mohlke KL, Silander K, Ally DS, Chines P, Blaschak-Harvan J, Douglas JA, Duren WL, Epstein MP, Fingerlin TE, Kaleta HS, Lange EM, Li C, McEachin RC, Stringham HM, Trager E, White PP, Balow J Jr, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kudelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB, Rha D, Segal L, Shapiro S, Sharaf R, Shurtleff B, So A, Tannenbaum J, Te C, Tovar J, Unni A, Welch C, Whiten R, Witt A, Kohtamäki K, Ernholm C, Eriksson J, Toivanen L, Vidgren G, Nylund SJ, Tuomilehto-Wolf E, Ross EH, Demirchyan E, Hagopian WA, Buchanan TA, Tuomilehto J, Bergman RN, Collins FS & Boehnke M 2000 The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics FUSION study. II. An autosomal genome scan for diabetes-related quantitative-trait loci. *American Journal of Human Genetics* **67** 1186–1200.

Weiss KM & Terwilliger JD 2000 How many diseases does it take to map a gene with SNPs? *Nature Genetics* **26** 151–157.

Williams RC, Knowler WC, Butler WJ , Pettitt DJ, Lisse JR, Bennett PH, Mann DL, Johnson AH & Terasaki PI 1981 HLA-A2 and Type 2 (insulin independent) diabetes mellitus in Pima Indians: an association of allele frequency with age. *Diabetologia* **21** 460–463.

Wiltshire S, Hattersley AT, Hitman GA, Walker M, Levy JC, Sampson M, O'Rahilly S , Frayling TM, Bell JI, Lathrop GM, Bennett A, Dhillon R, Fletcher C, Groves CJ, Jones E, Prestwich P, Simecek N, Subba Rao PV, Wishart M, Foxon R, Howell S, Smedley D, Cardon LR, Menzel S & McCarthy MI 2001 A

genome-wide scan for loci predisposing to type 2 diabetes in a UK population (The Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *American Journal of Human Genetics* **69** 553–569.

World Health Organisation Study Group 1985 Diabetes mellitus. *WHO Technical Report Series* no. 727.

Yamagata K, OdaI N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV, Chen X, Cox NJ, Oda Y, Yano H, Le Beau MM, Yamada S, Nishigori H, Takeda J, Fajans SS, Hattersley AT, Iwasaki N, Hansen T, Pedersen O, Polonsky KS, Turner RC, Velho G, Chèvre J-C, Froguel P & Bell GI 1996*a* Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young MODY3. *Nature* **384** 455–458.

Yamagata K, Furuta H, Oda N, Kasiaki PJ, Menzel S, Cox NJ, Fajans SS, Signorini S, Stoffel M & Bell GI 1996*b* Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young MODY1. *Nature* **384** 458–460.

Yeo GS, Farooqi IS, Aminian S, Halsall DJ, Stanhope RG & O'Rahilly S 1998 A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nature Genetics* **20** 111–112.

Zavattari P, Lampis R, Motzo C, Loddo M, Mulargia A, Whalen M, Maioli M, Angius E, Todd JA & Cucca F 2001 Conditional linkage disequilibrium analysis of a complex disease superlocus, IDDM1 in the HLA region, reveals the presence of independent modifying gene effects influencing the type 1 diabetes risk encoded by the major HLA-DQB1, -DRB1 disease loci. *Human Molecular Genetics* **10** 881–889.

Received 3 July 2001
Accepted 2 October 2001