Perspectives in Diabetes
Genetic Analysis of NIDDM
The Study of Quantitative Traits
Soumitra Ghosh and Nicholas J. Schork

Many studies are in progress worldwide to elucidate the genetics of NIDDM. Nevertheless, few articles are available that combine the interdisciplinary fields of medicine, genetics, physiology, and statistics in order to provide the scientific rationale for such an endeavor. Here we describe the methodology and background necessary to study the genetics of NIDDM and discuss how to analyze the data. We also provide a detailed bibliography for researchers and a glossary for those who are not experts in the field. In particular, we wish to emphasize the analysis of intermediate quantitative traits as a means to dissect the genetic basis of NIDDM. Diabetes 45:1-14, 1996

Following on the heels of recent successes with positional cloning efforts in the isolation of genes for rare monogenic diseases like cystic fibrosis (1), research is now focusing on the application of similar methods to dissect the genetic determinants of common complex diseases, such as hypertension, schizophrenia, coronary artery disease, allergic diathesis, IDDM, and NIDDM. By complex diseases, we mean diseases thought to be influenced by more than one gene or environmental factor, which hence do not exhibit a simple mode of inheritance (e.g., dominant, recessive). However, the successful extrapolation of positional cloning methodology for single gene disorders to the analysis of complex traits is not readily possible without more careful study designs and more sophisticated analytical and computational tools. The main reason for this is that genes influencing complex disease traits like hypertension and NIDDM generally induce more susceptibility to (as opposed to conferring strict determination of) the disorder and as such may not lead to disease unless other genes and/or environmental factors are also present. Therefore, unlike the one gene/one trait paradigm assumed for monogenic diseases, genes involved in a complex disorder may be neither necessary nor sufficient for disease expression. This poses considerable problems for classical linkage analysis–based gene-mapping strategies, which are more powerful at detecting loci that are necessary and in most cases sufficient for disease expression (2).

The format of this review is as follows. In the sections that follow, we review the evidence for a genetic component in NIDDM and critique traditional methods and approaches used to isolate genes involved in NIDDM. In subsequent sections, we describe a rationale for treating aspects of NIDDM as quantitative disorders and focus on study design and analytical issues relevant to quantitative trait analysis. The final section offers a brief summary and highlights areas for future research. We do not provide great detail on the physiology of diabetes or on positional cloning methods because the reader has great many options for these (3-11). An appendix is also provided for ease of understanding unfamiliar terms.

BACKGROUND
The evidence for a genetic component in NIDDM. While NIDDM is a heterogeneous disorder, it is characterized by a predominant phenotype in many societies (12). The environment has a strong role to play in NIDDM susceptibility, as shown by the less than 100% concordance rate for diabetes in identical twins (see below) and the significant contributing effects of diet and exercise (13). The prevalence of NIDDM varies markedly worldwide and even within a single nation. In the U.S., for example, there exist many ethnic groups, such as Mexican-Americans, African-Americans, and Pima Indians, with high frequency rates for NIDDM. Thus, the Pima Indians have a prevalence of ~35% after the age of 20 years as compared with the U.S. national figures of ~7% (14). The effects of racial differences are well illustrated in cases of genetic admixture between two populations with differing frequencies of disease. In one well-documented study, Knowler et al. (15) showed that the prevalence of NIDDM in those with mixed heritage in the Pima Indian community is inversely related to the extent of interbreeding with whites.

In addition, there are many cases where genes interact with environment to produce NIDDM. For example, it has been shown that among Asian Indians who migrate to Europe from the Indian subcontinent there is a markedly increased incidence of NIDDM as they become exposed to...
more Western lifestyles (16,17). Family and twin studies provide evidence for the genetic basis of NIDDM. However, twin studies are typically plagued by factors that may be difficult to account for, such as dissimilar intrauterine environments for each twin, knowledge of the true zygosity status, incomplete ascertainment, possibly inaccurate assumptions (e.g., that the environmental influences act in a similar manner for both dizygotic and monozygotic twins), and variable age of onset.

Monozygotic pairwise concordance rates in NIDDM have been estimated to be anywhere between 20% (29 of 142) (18) and 90% (48 of 53) (19). Reasons for the disparity include the different ascertainment criteria used for the individual studies. For a Finnish study (18), a population of twins was analyzed by record linkage but no clinical testing for NIDDM status was performed. Therefore, the actual percentage concordance rate may be much higher. In a British study (19), doctors were asked to report twins in which at least one twin was diabetic. Unfortunately, identical concordant twins were overrepresented, even though the diagnosis was clinically confirmed. A twin study that was both population-based and involved clinical testing was focused on unselected American army recruits who were followed up over 10 years (20). In this study, the pairwise concordance rate for monozygotic twins in white men was 41% (14 of 34). Furthermore, 65% (13 of 20) of nondiabetic monozygotic co-twins of diabetic probands had elevated blood glucose levels, suggesting that they were predisposed to diabetes. By 52–65 years of age, the estimated risk for monozygotic co-twins of a proband was six times more than the population prevalence.

Because of the age-dependent prevalence of NIDDM, it is to be expected that if twins were followed for a longer period of time, the concordance rate for monozygotic twins would be significantly higher than the figures presented here. Based on data from published twin studies, the age-adjusted monozygotic twin concordance rate may be as high as 70–80%. It is very important to note that in two of the above studies the concordance rate for monozygotic twins was significantly higher than for dizygotic twins (the dizygotic concordance rates were 9% [4 of 42] for the army recruit study [20] and 9% [27 of 307] for the Finnish study [18]; the British study [19] examined monozygotic twins and no dizygotic concordance rates were reported). This large difference between monozygotic and dizygotic twin concordance rates indicates the high degree of heritability for NIDDM (it is believed that the dizygotic twin rates are artificially low for much the same ascertainment reasons mentioned above).

Segregation analysis of NIDDM has, for the most part, excluded the hypothesis that NIDDM is controlled by a single major gene (12), although it has been shown that for a rare monogenic subphenotype of NIDDM, maturity-onset diabetes of the young (MODY), and in certain populations with a high prevalence of NIDDM (21,22) there may be a single major diabetes gene being transmitted. However, it is generally conceded that segregation analysis is less reliable for complex diseases (23). Risch (24) has developed methods to estimate the number of genes (under limiting assumptions) contributing to a complex trait, given the risk-to-relative information parameterized by \( \lambda \). This parameter \( \lambda \) is a ratio that compares the risk of disease to a relative of a proband to the population prevalence for the disease. This information is often accessible and less difficult to obtain than reliable multilocus segregation information. Nevertheless, using risk-to-relative information for NIDDM of the type suggested by Risch, it has been shown that a wide range of multifactorial models are compatible with NIDDM susceptibility (3).

**Candidate gene analyses in NIDDM: association studies.** Traditionally, the candidate gene approach has been the major strategy in attempts to identify genes influencing complex diseases, including NIDDM. With the candidate gene approach, the investigator assumes a priori that a specific gene that encodes a protein is involved in the pathogenesis of the disease. Variants of this gene are then tested for association with the disease in populations or for linkage with the disease in families. Most candidate genes are considered true "candidates" for biological reasons, such as when a gene is expressed in a relevant tissue or when an evolutionary homologue of a certain gene is found to influence disease susceptibility in another species. Possible candidate genes for NIDDM include genes coding for glucose transporters, the insulin receptor, insulin receptor protein kinase, intracellular protein kinases, and specific enzymes systems such as glycogen synthase and glucokinase (25), as discussed below. Muscle, fat cells, pancreatic \( \beta \)-cells, and the liver are the most probable sites of action for the deleterious products of the mutant forms of these genes.

Although association studies involving defined case-control samples have been the main analytical tool in the candidate gene approach, they actually seek to define or investigate one of two things: 1) possible linkage disequilibrium between alleles at a marker locus near a candidate gene and the putative disease or trait locus or 2) a direct (causal) association between a detectable polymorphism in the candidate gene and the trait or disease itself. Depending on the situation, it is important to realize that if a marker locus is near a disease locus there may be evidence for linkage in the absence of association (e.g., if the common allele of the marker locus is associated with the rare allele of the disease gene, which in general reduces the power to detect association) or association in the absence of linkage (for example when allele frequencies are high, as shown by studies of the insulin locus, which in many instances showed strong association but weak or no evidence for linkage to IDDM) (26).

The major problem underlying association analysis is the occurrence of spurious differences in allele frequencies between cases and control subjects. This may arise because the populations are not matched for factors such as ethnicity, age, obesity, or the extent of assortative mating (27). To help reduce the chance of a spurious association, an alternative approach to testing association in unrelated cases and control subjects can be used (28–31). This alternative method uses unaffected relatives of the cases (or "non-predisposing genotypes") as control subjects instead of unrelated individuals. Since NIDDM is a late-onset disease, probably the best control subjects to use are unaffected siblings (26) (because offspring may be too young to evaluate and parents may have died). There have been no studies to date in NIDDM that have used the intrafamily control method, although it has been successfully applied to IDDM (29).

An alternative to the use of control subjects is to test the frequency of transmission of the putative disease allele from heterozygous parents to affected offspring (29). Under Men-
nearly 250 other candidate genes tested so far for association with NIDDM may exacerbate or induce rare forms of NIDDM, they do not play a major role in common NIDDM because of some reports of an increase in maternal transmission of the disease (38). An individual's mitochondrial genome is primarily transmitted by the mother (38), and in addition it has been proposed that naturally occurring age-related mutations in the mitochondrial genome could lead to a decrease in capacity for oxidative phosphorylation with a consequent reduction in insulin secretion from the islets (35–37). Alternatively, apparent maternal transmission could reflect environmental (in utero) risk due to uterine exposure. In summary, although mutations in the mitochondrial genome may exacerbate or induce rare forms of NIDDM, they seem unlikely to play a major role in common NIDDM (39).

In addition to the genes discussed above, there have been nearly 250 other candidate genes tested so far for association or linkage to NIDDM, but none have shown consistently significant results across different study populations (7,40). However, in most cases it is unclear whether better study designs and larger sample sizes would have produced different results. MODY serves as a good example of candidate gene analysis applied to diabetes research. Glucokinase gene mutations account for >50% of all cases of MODY and associated impaired glucose tolerance (IGT) in French pedigrees (41), although MODY in the RW kindred of Michigan is mainly linked to adenosine deaminase on chromosome 20q (indicating genetic heterogeneity) (42). Here adenosine deaminase represents simply a marker and not the contributing locus. A third mutation on chromosome 12q accounts for approximately half of all unresolved cases of MODY in France and is associated with a severe insulin secretory defect (43). Diabetes in most French MODY pedigrees with the glucokinase mutations is thought to arise from an increased glucose threshold for insulin secretion and also from impaired hepatic uptake and metabolism of glucose (44–45). Despite the fact that some populations show a significant association of NIDDM with glucokinase polymorphisms, glucokinase mutations per se are unlikely to explain more than 1% of all NIDDM cases (7,46–55).

Linkage analysis and positional cloning for multigenic disorders. As the number of biologically relevant candidate genes in NIDDM becomes too numerous to study simultaneously, a positional cloning approach (where no a priori hypothesis is made as to the type or location of genetic mutations in NIDDM) becomes a more attractive alternative strategy. Principles of positional cloning for monogenic disorders are described in detail by Collins (6). A more recent article (56) describes the changing technology and the emphasis on a positional candidate approach. Here the combination of linkage information and physical and transcript map data is used to home in on the gene. However, as mentioned above, there are problems in applying positional cloning techniques to complex diseases. The strict one-to-one relationship between genotypes and phenotypes in monogenic disorders breaks down for complex diseases. The main reasons for this are that multiple genes more than likely influence most complex diseases, which, in turn, makes it difficult to isolate and characterize the effects of each and every individual locus determining the disease. Such multiple genes may interact (i.e., exhibit epistasis) or induce susceptibility independently (i.e., exhibit locus heterogeneity). In addition, most complex diseases like NIDDM are influenced by often unknown environmental factors as well. As a consequence, fine genetic mapping at the resolution possible for a monogenic disorder will be difficult to achieve in a complex disorder simply because the effect of any one gene may be low relative to the combined impact of the others (e.g., the penetrance associated with any one gene may be very low). Furthermore, epidemiological studies for a complex disease will also be difficult because some diabetic subjects will not have the mutation in a positionally cloned gene (i.e., locus heterogeneity), while many unaffected individuals (i.e., reduced penetrance). This illustrates the profound problem of identifying disease pathways that can be closely correlated with genotype abnormalities.

Although the precise mapping of genes influencing complex traits may be extremely difficult given the problems alluded to above, it should now be possible to determine the approximate location of genes with large to moderate effect on complex disease expression given the availability of an abundant supply of genetic markers and, in particular, of microsatellite markers. Available markers have set the resolution of the current human genetic map today in the neighborhood of 1 cM (centimorgan) (57–59). The higher resolution of the genetic map will make it easier to map genes. This is due in part to the simple fact that a denser map (more closely placed markers) is likely to yield a marker nearer to the disease locus than a sparser map.

Calculations for the genetic analysis of IDDM suggest that a 10-cM resolution map with 96 affected sib-pairs will suffice to map loci having a λc (Risch's defined risk to a sibling of an affected person compared to the population prevalence) of 2.5 with 99.9% probability (2.5 is the approximate effect of the human leukocyte antigen [HLA] in IDDM). Thus, Davies et al. (60) were able to map two new loci, IDDM4 and IDDM5, to chromosomes 11q and 6q, respectively, using a sib-pair strategy that made use of a genome screen with
microsatellite markers spaced roughly every 11 cM. The \( \lambda_s \) for IDDM4 was given as 1.2, suggesting that the Davies et al. result for IDDM was somewhat unexpected given the theoretical assessment of the power of the method they used. A separate calculation showed that to have a good chance to detect all loci with \( \lambda_s \) between 1.2 and 1.8 would require several hundred sib-pairs and a 3-cM genetic map (60).

As with association studies, the question of significance levels for declaring linkage in a complex disease with multiple loci and the corrections required because of multiple testing need to be addressed. Issues related to the statistical significance levels for genome-wide scans with multiple DNA markers are discussed by Feingold and colleagues (62,63) and given by Lander and Schork (64). As an example, if parents are unavailable for study (a safe assumption for NIDDM) and one has to rely on an affected sib-pair strategy, a logarithm of odds (LOD) score of at least 4 (which would correspond to a single-locus \( P \) value of \( 2 \times 10^{-5} \) or less) would be required to achieve an overall false-positive linkage rate of 0.05 and reasonable power. Nevertheless, it would be prudent to follow up any \( P \) value that is 0.01 or less at the first-pass genome-wide screening with more markers and families to exclude any false-positives. In addition, it is important to give \( P \) values for linkage that are a reflection of the sample size and not an asymptotic result. Thus, \( P \) values based on simulation are now encouraged, and statistical packages are now being developed to compute these (J. Buhler et al., personal communication).

Statistical tools for gene mapping involving complex traits should make minimal assumptions because one will rarely know a priori how many genes will be involved in the disease and what impact these genes might have on disease susceptibility. Therefore, model-independent approaches (nonparametric methods) are of primary importance in the analysis of a disease locus. Although standard linkage analysis can be relatively impervious to model misspecification in certain situations (65) and in most cases more powerful than nonparametric methods, it is advisable to consider both approaches in any one analysis. Two model-independent methods are commonly used in human genetics and make use of affected relatives given gene identity-by-descent (IBD) and identity-by-state (IBS) information. Unaffected persons provide no definite information for complex diseases, and the use of unaffected/affected relative pairs is a less powerful approach than the use of affected relative pairs in most cases (66,67). However, the degree of information gained by analyzing unaffected/affected pairs depends on both the recurrence risk to the relative concerned (less powerful with values of <0.5) and, in the case of quantitative traits, whether the trait values are in the extremes of the distribution. In addition, genetic data from the unaffected person can give increased phase information and can enhance genotype error detection rates.

In IBD-based methods, parental genotype information is required (but in some cases this may be inferred). The aim of these methods is to look for skewing in the distribution of transmitted alleles from Mendelian expectations in affected relatives (Fig. 1). That is, affected sib-pairs at an unlinked locus share 0, 1, or 2 genes IBD at proportions 0.25:0.5:0.25, respectively (66). A significant distortion from these expected frequencies, with an increase in siblings sharing zero alleles IBD and a decrease in siblings sharing zero alleles IBD, suggests that the marker is linked to (or possibly associated with) a disease locus. Many tests are available to measure deviations in allele-sharing proportions from expected proportions, and recently these tests have been made more powerful (68,69).

IBS approaches are not concerned with the transmission of alleles but rather rely on the simple identity of alleles among related individuals. Thus, persons with genotypes AB and AC share, by state, the A allele, although it may be the case that the A alleles possessed by each are not IBD (i.e., transmitted to them both through a common ancestor who possessed an A allele [Fig. 1]). IBS methods are useful in cases of late-onset disease (like NIDDM) where parental data are usually not available. IBS methods can also be used in situations where markers are not highly polymorphic (68,70) and become almost as powerful as IBD methods when markers are fully informative in the sample. Recently, IBS methods have been modified to include information from all relative pairs (71) and extended to multiple linked markers (72). It is essential to incorporate allele frequency information into IBS analysis methods, since more common alleles will be shared by affected individuals simply because they are more common and not because they are coupled or associated with disease alleles.

At least five factors influence the power to detect linkage involving a complex disease: 1) \( \lambda_s \), the risk to a relative of an affected individual given that they share a certain susceptible allele or genotype; 2) \( \theta \), the presence of any interaction between multiple genes; 3) \( \theta \), the recombination fraction between the susceptibility and marker loci; 4) the heterozygosity (i.e., informativity) of the marker(s) used; and 5) the relationships of relatives studied (67,70). Theoretical calculations by Risch (67) suggest that to detect a single locus that has \( \lambda_s \) of, say, 1.6 with 90% power in NIDDM (for which the approximate total age-unadjusted \( \lambda_s \) is estimated to be 3.5 [3]), use of nonparametric sib-pair linkage test would require 300 affected sib-pairs (ASPs), provided that the heterozygosity of the markers is 1 and the map consists of markers at least every centimorgan. If \( \theta \) is assumed to be 0.05, then the power drops to 70% with 300 sib-pairs (67). Clearly, >300 sib-pairs will be required to detect linkage for NIDDM, since markers are rarely fully informative or equally spaced and

**FIG. 1.** Graphical representation of situations in which IBD and IBS information can be computed explicitly or is ambiguous. The numbers represent alleles (genotypes) at a single locus. A: the two offspring share no alleles IBS and no alleles IBD. B: the two offspring share one allele IBS and the same allele IBD. C: the two offspring share both alleles IBS and IBD. D: the two offspring share both alleles IBS, but only one allele is unambiguously shared IBD (it is unclear whether the two offspring were transmitted the same allele from the father). E,F: the two offspring share two alleles IBS, but it is unclear how many alleles they share IBD. Adapted from Fig. 4 in Schork and Chakravarti (67a).
there is likely to be genetic and environmental interaction in NIDDM. However, when non–fully informative markers exist, multipoint linkage analysis (i.e., use of more than a single genetic marker in the analysis) can be used to increase the power of detecting linkage and locating a disease susceptibility gene. This has been shown for both model-dependent (73) and model-independent methods (70,72,74–76). Multipoint or interval mapping methods should always be used to refine the region of interest in positional cloning strategies.

Once linkage to a DNA marker has been established, polymorphic marker data can also be used to study linkage disequilibrium between loci that are physically close to the disease locus (say a few hundred kilobases or nearer). This strategy exploits the knowledge that affected chromosomes descended from a common ancestral chromosome on which the mutation arose should show a distinctive haplotype in the immediate vicinity of the mutation, which would reflect the haplotype of the ancestral (i.e., original, mutant) chromosome. A special application of disequilibrium mapping of this sort based on a branching processes model of mutant haplotype transmission in an isolated Finnish population revealed a locus for a rare monogenic disorder known as diastrophic dysplasia (77). Such isolated homogenous subpopulations of the type observed in Finland are unlikely to be found in the largely outbred, highly admixed U.S. population. The disequilibrium mapping method used to map the diastrophic dysplasia gene was able to narrow the interval known to contain the diastrophic dysplasia gene from 1.5 cM (the limit of linkage mapping) to 60 kb (77). Linkage disequilibrium mapping of this sort may prove useful in fine mapping for complex disease genes, and this will be especially true for relatively homogenous populations segregating for one or a few recent mutations at separate loci, such as Finland.

Interestingly, not all regions of the genome show a monotone relationship between distance and extent of disequilibria between markers (78). Furthermore, the power to detect linkage disequilibrium is largely dependent on marker allele frequencies (27). Hill and Weir (79) and more recently Terwilliger (80) have proposed likelihood-based methods to estimate gene location parameters through disequilibrium. Hill and Weir (79) conclude that for intermarker distances below 40–50 kb, linkage disequilibrium will not, in general, correlate well with physical distance in most populations. The relationship is more exact, however, at greater distances (up to ~70 kb). This result is corroborated by Watkins et al. (81). Design and sample-size considerations for linkage disequilibrium studies have recently been given by both Olson and Wijsman (82) and Terwilliger (80). Although the figures reported by these authors are for simple Mendelian disorders, it is instructive to note the significant loss in power for most models when parental haplotypes are unavailable.

Phenotypes. Because there are likely to be a number of genetic and environmental risk factors for NIDDM, which may vary both within and between populations, it may prove to be difficult to identify a single universal pathogenic mechanism responsible for NIDDM susceptibility. This has obvious implications for the genetic analysis of NIDDM, as suggested earlier. By studying a more carefully selected ethnically and clinically homogenous subgroup, it may be possible to reduce much of the complexity and etiological heterogeneity for NIDDM in a given sample of patients. For example, MODY is a clear subtype, showing an earlier age of onset with a major gene controlling the expression of the trait. However, to be useful at all, such subtypes must be heritable and not so rare as to make difficult the appropriate sampling of probands.

One of the most important aspects of genetic analysis is accurate phenotyping for the traits to be studied. Insulin resistance, for example, is affected by age, diet, exercise, stress, medication, and hormonal status of the individual. These factors should be standardized as much as possible before measurement of insulin resistance, although it may also be possible to statistically adjust for such factors during the analysis of the data. Furthermore, phenotypic misclassification in genetic studies will generally bias the recombination fraction estimate upwards and reduce evidence for linkage (65,83). As an example, gestational diabetic subjects, who at the postpartum stage have normal glucose tolerance, or previously obese individuals with IGT who now have normal glucose tolerance and normal body weight may present problems in phenotype classification. However, it is entirely possible to analyze these individuals as a separate class even though the sample sizes will be small. In light of these and related issues, Alberti (84) created a classification for individuals with IGT (who have an increased but highly variable risk of becoming diabetic). He concluded that IGT represents at least four types of individuals: 1) diabetic subjects who were low on the day of testing, 2) normal subjects who happened to have high values, 3) individuals on the way to developing diabetes, and 4) those stable in the IGT category.

The diagnosis of diabetes also poses problems. Recently, the fasting plasma glucose level (as defined by the World Health Organization [85] and the National Diabetes Data Group [86]) has come under scrutiny as a method for the accurate diagnosis of diabetes compared to the standard oral glucose tolerance test (OGTT) (87). As defined, the sensitivity of a fasting plasma glucose of >7.8 mmol/l (for diabetes) can range in different population groups from 22 to 91% and is most commonly <35%. However, the standard OGTT is also prone to variation. The amount and type of glucose used, body muscle mass, extracellular volumes, and gastric absorption will all make a substantial difference in the metabolic profile used to assess the OGTT. In contrast, a strong case has been made to gather information on insulin secretion and resistance from OGTT data (88). Obviously, it is important to appreciate that no amount of sophisticated statistical analysis can ever make up for inaccurate or less than excellent clinical or biochemical phenotyping.

**ANALYSIS OF QUANTITATIVE TRAITS IN NIDDM**

In quantitative trait locus (QTL) analysis, categorical diagnostic states are not required; all that is needed is a continuously measurable observation such as $S_i$ (which is a measure of insulin sensitivity, the inverse of insulin resistance). For quantitative traits, it is assumed that observations represent a mixture from two or more groups of individuals, defined by the genotypes that they carry at a locus. Thus, for example, individuals with AA genotypes may have higher values of a trait than those with Aa or aa genotypes. Of course, mixing of this sort is not often obvious from phenotype information alone. With this in mind, it should be clear that the greater the overlap of the component distributions in
the mixed distribution, the harder it is to distinguish underlying genotypes (89). Power to detect linkage in quantitative trait analysis depends on at least five factors: 1) the distance between mean phenotype values associated with underlying genotypes, 2) the underlying allele frequencies, 3) the density of marker coverage, 4) heterozygosity of the markers used, and 5) the sampling scheme used to collect data.

Genetic analysis of NIDDM has primarily concentrated until now on the analysis of discrete diagnostic categories. Given the difficulties in the classification of discrete traits like NIDDM alluded to above, this form of analysis might benefit by being complemented with the study of other important heritable quantitative traits such as $S_B$ which is known to both predate and predict NIDDM. Intermediate quantitative traits are more proximal to the mutated genotypes in terms of pathogenesis compared to NIDDM per se and may be controlled by fewer loci and environmental factors. If this were true, then it would be easier to dissect the separate major genetic contributions to NIDDM by intermediate phenotype analysis as opposed to discrete trait analysis (90). In conclusion, for QTL analysis to be successful, such traits must be influenced by or influence a limited number of pathogenic processes leading to NIDDM (to prevent factors that might obscure the detection of loci), be largely heritable, and be predictive of (or at least related to) the final disease outcome (NIDDM). Advances in computing power to detect factors that might obscure the detection of loci), be largely heritable, and be predictive of (or at least related to) the final disease outcome (NIDDM). Advances in computing and statistical methodology, the fact that accurate and reproducible quantitative trait measurements can be performed, and the availability of dense genetic maps with highly polymorphic markers make it reasonable to attempt to map human QTL at the present time.

**NIDDM quantitative genetics: lessons from glucose studies.** The full spectrum of normoglycemia to IGT to diabetes can be assessed not only in terms of affection status but more naturally as a continuum. Not surprisingly, the only trait linked to NIDDM status that has been thoroughly investigated quantitatively is blood glucose level. There is no strong consensus as to whether a single genetic locus controls blood glucose levels in humans. Evidence from mice strains does, however, point to polygenic control for this trait (91). There have been a wide variety of study designs used to investigate glucose levels in humans, in terms of both study populations and analytical methods.

For instance, Beaty and Fajans (92) found evidence for additive alleles at an autosomal locus that accounts for $\approx 25\%$ of the variance of age-adjusted log fasting blood glucose levels in families that were ascertained via probands with NIDDM. This included 8 of 337 individuals with fasting plasma glucose $>160$ mg/dl. Elston et al (21) studied two separate populations of Seminole Indians ascertained on the basis of health surveys and used mixture distribution models to study logarithmic 1-h post-glucose load measurements. No exclusions were made in this study for diabetic individuals. In only one of these populations did two equally likely but different genetic models (dominant or recessive) appear to be compatible with the data. Boehnke et al. (93) studied adjusted fasting blood glucose levels in normoglycemic three-generation white pedigrees using a variance components approach (94) and found that polygenic loci could account for as much as 27% of the variability of the trait. In one additional study, segregation and path analysis applied to both Japanese-Americans and an Israeli population failed to show significant genetic determinants for fasting blood glucose (95,96).

One of the reasons why the studies of blood glucose have so far been inconsistent may be that the traits do not have all the properties for good QTL analysis as described in the previous section.

**Other quantitative phenotypes in NIDDM.** Important quantitative trait measurements of relevance to NIDDM can be derived from glucose clamp studies or from frequently sampled intravenous glucose tolerance tests (FSIGTs) using the minimal model. The glucose clamp techniques are used for quantifying insulin secretion and insulin resistance. The glucose infusion to maintain steady state in the clamp technique is an indication of insulin resistance, whereas plasma insulin responses (for the hyperglycemic clamp) reflect pancreatic function (97). Alternatively, the minimal mathematical model for glucose utilization gives estimates for insulin resistance and pancreatic responsiveness using glucose and insulin levels from a FSIGT using either tolbutamide or insulin injections (10). This approach does not require any assumptions on steady-state glucose or insulin levels. A nonlinear least-squares procedure from data gathered from the FSIGT gives estimates for glucose effectiveness ($S_G$), insulin sensitivity ($S_i$), and first phase ($\varphi_1$) and second phase ($\varphi_2$) insulin secretion. Insulin resistance and low $S_i$ refer to the same physiological phenomenon. The minimal-model method can be adapted for population-based studies and has reasonable reproducibility (38) and thus offers advantages over the traditional glucose clamp approach (10). Because the estimates for insulin resistance and glucose effectiveness from the tolbutamide-modified minimal model and glucose clamp are highly correlated at 0.7–0.9 (99,101), it seems reasonable to use the minimal model in large-scale studies.

One potential drawback is that because diabetic individuals lose their first-phase insulin response, one cannot measure their insulin resistance with the minimal model and tolbutamide. In addition, diabetic individuals have a strong acquired insulin resistance independent of etiology, which makes it difficult to understand the significance of a single quantitative measure. Another recent finding is that the minimal model overestimates $S_i$ and hence underestimates $S_I$ (102). These differences are due to the fact that the minimal model assumes a single compartment for glucose. However, the correlation between the minimal model values and those from the correct two-compartment model are very high, allowing meaningful interpretation when using $S_I$ and $S_0$ for genetic analysis (R. Bergman, personal communication). It is possible to give exogenous insulin to diabetic subjects to estimate their $S_i$ and a similar approach can be used to obtain an $S_I$ value from a nondiabetic subject. For these cases, the correlations between the glucose clamp and insulin-modified minimal model values are not as strong ($-0.55$) (103).

One of the most important findings in recent years is that physiological variables such as $S_I$ and $S_0$ are lower in NIDDM-susceptible individuals than in nonsusceptible individuals. Martin et al. (104) showed that simultaneous occurrence of both low $S_I$ and low $S_0$ (i.e., below median values) was a good predictor of future NIDDM, with low $S_I$/high $S_0$ and high $S_I$/low $S_0$ showing intermediate risks and high $S_I$/high $S_0$ showing the least risk. These abnormalities were apparent many years before the diagnosis of NIDDM was
made. It was of related interest that $p = 1$ was not decreased in the prediabetic stage, although it did exhibit familial aggregation. Therefore, changes in this variable were not predictive of diabetes. The study by Martin et al. (104) can be criticized because it investigated 155 people who were offspring of two NIDDM patients and were followed up over 6–25 years. Thus, they were probably not representative of NIDDM populations at large. However, in support of the conjecture that low $S_1$ predicts diabetes, McCance et al. (105) also found in a Pima Indian population that a history of parental diabetes was strongly correlated with the subsequent development of diabetes in offspring. Two-hour plasma glucose, fasting serum insulin, and relative weight (observed weight compared with expected) were the significant predictors for diabetes for those with one diabetic parent. The same conclusions were drawn for a random group of nondiabetic Pima Indians (106). The strongest predictor for future NIDDM in two of these three studies (104, 106) was insulin resistance (low $S_1$) independent of obesity, sex, or age. $S_0$ and $S_1$ have been found to be uncorrelated, and $S_1$ shows evidence for familial clustering (107). In addition, evidence for a major gene with a dominant mode of inheritance contributing to $S_1$ in a subgroup of patients with very low overall $S_1$ values has been found (107).

Other investigators have observed that measures of insulin resistance such as $S_1$ (107), fasting insulin levels (108), and MaxM (maximal insulin action, which is a measure of maximal insulin stimulation of glucose uptake) (107) are heritable and possibly controlled by a single gene locus. For instance, Bogardus et al. (22) have shown that the distribution of MaxM is characterized by a mixture of normal distributions, which is suggestive of a major genetic effect. In addition, Schumaker et al. (108) have used segregation analysis to show that fasting insulin is under genetic control. The latter study suggests the presence of a major gene for insulin resistance that is recessive and accounts for 33.1% of the total phenotypic variance with an allele frequency estimated to be 0.25. The presence of a major locus for insulin resistance has thus been confirmed in at least two separate populations, in Pima Indians and in whites in Utah. Still further and more recent evidence comes from a study on Mexican-Americans in which log-transformed 2-h insulin levels segregated as a monogenic trait (110). In addition, these studies have suggested that the genetic effect determining insulin resistance is independent of obesity, age, or sex. Therefore, with the use of standard likelihood-based linkage analysis and with large samples, it will be feasible to map a locus for insulin resistance.

Linkage studies of insulin resistance corroborate the suggestion that a major gene may contribute toward this trait. For instance, the gene $FABP2$ (fatty acid–binding protein 2) on chromosome 4q shows evidence for linkage to MaxM in Pima Indians, and there is unpublished evidence that this gene is actually implicated (M. Prochazka, personal communication) (111). This study used sib-pair linkage analysis on nondiabetic subjects measured for MaxM. Diabetic individuals were not included in this study because of the secondary changes in insulin action attributed to the diabetic state itself. $FABP2$ is a candidate gene for NIDDM because the protein it encodes determines the composition of fatty acid uptake from the intestine, which could alter glucose dynamics in muscle tissue, leading to insulin resistance. To assess the relationship between $FABP2$ polymorphisms and glucose tolerance further, a European consortium that took its lead from the Pima Indian study looked at two traits, insulin sensitivity and diabetes, in three different white European populations (112). Log-linear analysis of the results confirmed a weak statistical interaction between glucose tolerance (as defined by diabetic status) and the specific $FABP2$ allele (partial $\chi^2$ 14.4, df 6, $P = 0.027$). This interaction parameter differed significantly from 0 ($P = 0.046$).

Another quantitative trait influencing NIDDM susceptibility can be derived from evaluation of $B$-cell function by measuring the acute insulin response to glucose (A1Rglucose) (or to arginine) (25). This may be calculated as the mean increment above basal insulin levels measured at 2, 3, 4, 5, 6, 8, and 10 min after a maximally stimulating intravenous glucose bolus. This trait is familial and a possible positive predictor of diabetes (106, 113). Other quantitative traits that have been studied in less detail include insulin secretion and proinsulin levels. Insulin secretion is both disordered and irregularly pulsatile in prediabetic and diabetic subjects, which is not the case for nondiabetic subjects. Time-series analysis of such variations may provide a way of defining suitable quantitative traits for subsequent genetic analysis, though this will be very difficult to achieve on hundreds of individuals (114, 115). A further quantitative trait with potential for genetic analysis is proinsulin level. The ratio of proinsulin to insulin levels rises in NIDDM, but its variability in the prediabetic state may make it an unlikely predictor for future NIDDM (116, 117).

**Analytical methods used for quantitative traits.** Appropriate statistical tools are of extreme importance in the analysis of quantitative traits involved in NIDDM susceptibility. It is then perhaps understandable that the development of statistical tools appropriate for genetic analysis closely followed the debate about the nature of inheritance. At the beginning of this century, soon after Mendel’s work, a rift developed between two opposing schools of thought in genetics. On one side there were the “Mendelians,” who believed in the particulate nature of Mendelian inheritance (i.e., single segregating genes were responsible for most of the variation in phenotypes). On the other side were the “biometricians,” who noted that most traits in humans were quantitative in nature and therefore amenable to correlation and regression analysis. Fisher (118) brought the two schools together by postulating that the genetic contribution toward a quantitative trait was made up of many small independent additive genetic effects each controlled by a genetic locus (i.e., polygenic model). We now know that oligogenic models (which posit a few loci) are probably more characteristic of complex disease pathogenesis and that epistasis between such loci may play a major role in some complex traits (119, 120).

Before the mid-1970s, simple analysis of variance or regression methods (regressing phenotype on genotype) were mostly used to estimate linkage to quantitative traits with a single marker (but see Lange et al. [121]). An advance in quantitative trait analysis was made by Lander and Botstein (122), who introduced interval mapping for the genetic analysis of experimental populations. Interval mapping is more powerful than methods that work with simply one marker locus, since it tests the hypothesis that a disease locus is between two markers, as opposed to being unlinked to both. Use of interval mapping also allowed one to distinguish between a marker tightly linked to a QTL of small
TABLE 1
Some linkage analysis methods for human quantitative traits

<table>
<thead>
<tr>
<th>Basic method</th>
<th>Description</th>
<th>References</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Correlation model</td>
<td>Correlation between quantitative measures in siblings is used to investigate linkage.</td>
<td>Penrose (142)</td>
<td>Not implemented</td>
</tr>
<tr>
<td>2. Simple linear model</td>
<td>Regression of the squared difference in sibling (or arbitrary relative) pair trait values on the proportion of alleles shared IBD at a marker.</td>
<td>Haseman and Elston (127)</td>
<td>Not very powerful or flexible. Used often and simple to implement and interpret. Nonparametric version cannot distinguish between loose linkage and small effect.</td>
</tr>
<tr>
<td>4. Simple linear model</td>
<td>Nonparametric extension of the method of Hill.</td>
<td>Smith (144), Bener et al. (145)</td>
<td>Not powerful or implemented but nonparametric.</td>
</tr>
<tr>
<td>5. Simple linear model</td>
<td>Extension of Smith (144) combining information from families of different sizes and structures.</td>
<td>Cockerham and Weir (146)</td>
<td>Not very powerful, untested.</td>
</tr>
<tr>
<td>6. Standard pedigree analysis</td>
<td>Traces the cosegregation and recombination phenomena of marker alleles with putative trait-influencing alleles within pedigrees.</td>
<td>Ott (89), Lange et al. (121), Elston et al. (147)</td>
<td>Difficult computationally. Not easy to accommodate residual genetic or background environmental (e.g., covariate) effects. Many parameters must be estimated.</td>
</tr>
<tr>
<td>7. Regressive-model-based pedigree analysis</td>
<td>Modification of standard pedigree analysis approach that can model residual and/or covariate effects.</td>
<td>Bonney et al. (148)</td>
<td>Flexible. Not tested for oligogenic models. Must make assumptions about residual correlations.</td>
</tr>
<tr>
<td>8. Simulation-based estimation of pedigree models</td>
<td>Standard pedigree analysis approach with modern simulation-based procedures used to estimate relevant quantities.</td>
<td>Guo and Thompson (149)</td>
<td>Can be computationally difficult. May be difficult to interpret with multiple markers and assumed trait loci.</td>
</tr>
<tr>
<td>9. Variance component models</td>
<td>Partitions the variation of a quantitative trait into components, one of which effects degree of genetic similarity in terms of IBD alleles at a marker.</td>
<td>Amos (150), Schork (151)</td>
<td>Extremely flexible, but not powerful without selected samples. Untested.</td>
</tr>
<tr>
<td>10. Genomic region sharing variance components models</td>
<td>Modification of standard variance components approach but uses estimates of IBD sharing from multiple markers.</td>
<td>Goldgar (74), Schork (151)</td>
<td>Flexible but not necessarily powerful, untested.</td>
</tr>
<tr>
<td>11. Multi-marker linear model</td>
<td>Regression-based approach that considers regressing differences in trait values on IBD sharing among relatives based on multiple markers in a region of interest.</td>
<td>Fulker and Cardon (128)</td>
<td>More powerful than Haseman and Elston's (127) method. Can be made more powerful with selected samples. Not flexible (not designed to accommodate covariates).</td>
</tr>
<tr>
<td>12. Discordant sib-pair contrast</td>
<td>Investigates lack of allele sharing at a putative QTL between siblings with trait values at opposite ends of the trait distribution.</td>
<td>Risch and Zhang (152)</td>
<td>Awkward and potentially expensive sampling strategy, but powerful. Could be plagued by heterogeneity problems. Untested.</td>
</tr>
</tbody>
</table>

Effect and a marker loosely linked to a QTL of large effect. Furthermore, Lander and Botstein (122) defined significance-level based thresholds for relevant test statistics that can be used to declare linkage in the presence of multiple testing and discrete chromosome length and showed that by selecting progeny with extreme phenotypes it is possible to more efficiently map relevant genes. Selecting samples based on the extremes of a quantitative trait distribution has also been shown to increase power to detect loci that account for only 10-20% of the total phenotypic variance in human pedigrees (123). Multiple regression methods have also recently been developed for analyzing many disease loci together and have been further extended to human analysis settings, where unknown phase and multiple alleles make likelihood calculations more cumbersome (124). Finally, approaches for combining interval mapping and multiple regression techniques have been developed that can be used to resolve closely linked QTLs as well as to enhance precision mapping of a single QTL by conditioning on information from markers that are linked to other QTLs throughout the genome (125,126).

There are two broad approaches to human QTL analyses: analyses based on phenotype-on-genotype regression (or linear model analyses) and standard pedigree analyses with...
modifications appropriate for continuous expression of the phenotype. There are many derivatives of these two basic approaches (Table 1), some of which are more powerful and more widely used than others. Yet another method involves “measured genotype” information for quantitative phenotypes, which is primarily concerned with investigating the contribution of a known or specific allelic variability at a disease locus (129). The measured genotype approach assumes one has a candidate gene, and it will not be discussed in more detail here. Finally, it is noted that mapping QTLs in animals is an expedient way to potentially identify NIDDM susceptibility genes, as shown by the recent cloning of the ob gene (130).

**Multiple contributing loci and multiple traits.** A complicating factor in the analysis of any complex trait, whether qualitative or quantitative, concerns multiple locus effects. It seems clear that the “noise” induced by the segregation and effects of other loci may wash out or reduce the “signal” given by genes whose locations and effects are being sought. This will be a likely problem in NIDDM. Analytical methods for multiple loci detection must proceed from some sort of simultaneous search for multiple genes. Simultaneous search involves studying several candidate loci together to see whether they account for transmission of the trait, and it has been used in the analysis in experimental species only to date (122). From a mathematical point of view, the likelihood of the hypothesis that in each family the trait cosegregates with one of the loci is compared with the likelihood of the hypothesis that the trait is unlinked to any of these loci. Subsequently, all pairs of loci and then all triplets (and so on) are examined to find a set that together explain transmission in all families. This method needs more development with power and sample-size studies, incorporation of epistasis, and extension to more complicated multigenic models. A recent attempt has been made in studying relevant phenomena within one model (131).

Schork et al. (132) and Knapp et al. (133) have investigated two-trait-locus two-marker-locus linkage analysis in human genetics and have shown that in general more information is gained by analyzing loci together. However, at least with the former approach, the investigator still has to define some model parameters such as penetrance and gene frequency. In addition, the method is computationally intensive to the point of being prohibitive until further algorithmic and computational advances are made. More recently, Knapp et al. (134) have investigated a number of two-locus IBD ASP models in settings similar to those studied by Schork et al. (132). They found that the two-locus mean test proved to be the most powerful for the majority of underlying epistatic and heterogeneity models. The analysis was computationally feasible and again emphasized the increased power when analyzing more than one locus simultaneously. It seems that multi-locus analysis will be a regular part of any complex disease total genome scan, though not as an initial screening strategy (132,133).

In addition to the analysis of multiple loci, work considering the analysis of multiple phenotypes is extremely relevant to the genetic analysis of NIDDM. For example, in the context of NIDDM, it may be more powerful to map a major locus for both $S_p$ and acute insulin response to glucose simultaneously rather than mapping either trait on its own. This is especially true if the traits are shown to be strongly correlated. Most methodological work with multivariate traits so far is either untested (134,135) or restricted to the fitting of mixtures of bivariate distributions to the observed data (136). Lange and Boehnke (137) examined covariance component models with extensions to more than one trait. The Haseman and Elston (127) method has also been extended to include multivariate traits again assuming the pleiotropic activity of a few genes (138). This is an area, as in the analysis of multiple loci, that deserves greater attention.

**GENERAL PROBLEMS AND FUTURE STUDIES**

NIDDM is a common disease, but little is known about its genetic basis. Multiplex pedigrees and NIDDM sib-pairs with living parents are uncommon, which makes ascertainment of appropriate samples difficult (139). A nationwide television campaign may be more fruitful in gathering families (as proved in France) than simple clinic or record linkage-based ascertainment (140). In addition, other complicating factors may confound the detection of NIDDM susceptibility genes. For example, bilineality (the introduction of disease genes by multiple individuals in a single pedigree) was present in one study in 38% (11 of 29) of families ascertained through affected probands (139). Although recent work shows that bilineality in linkage analysis may not waste as much information as previously thought, it is probably still worth avoiding if at all possible (132,141). Sib-pairs are the most commonly available affected pairs but are not necessarily the most efficient in terms of genetic information (67). However, for small gene effects (as in NIDDM) sib-pairs are probably the best.

Collaborative studies may be needed to pool families (138,140) and to provide the necessary expertise in molecular genetics, statistical genetics, clinical physiology, and epidemiology, which is essential for a successful outcome. Countries with relatively isolated populations with possible founder effects, such as Finland, are going to be important sources for family material. Also of great importance is the question of reproducibility of a significant result in a separate population. When this is confirmed, the primary result gathers much more scientific credence. Cross-species investigations to look for conservation between genes and also to understand the underlying biology are going to be crucial subsequent steps.

Mapping NIDDM genes has rapidly become a major focus of several groups worldwide. In this context, quantitative trait analysis may also prove useful, but only if the chosen traits in NIDDM are largely heritable, are easy to collect, and have a strong biological relationship to the endpoints of diagnosis. Furthermore, more theoretical work is needed in the development of simultaneous search methods and in the analysis of multivariate traits. The difficulty in moving from the identification of a large region contributing to NIDDM susceptibility to the actual identification of the responsible gene should not be underestimated. Nevertheless, in the long run, it can be hoped that the genetics will help unravel the biochemical pathways leading to NIDDM as well as give us clues from epidemiological studies as to which environmental triggers are necessary for disease expression. These triggers may well depend on the specific constellation of genetic mutations in a given individual. Careful study design and statistical analysis should reduce the eventual size of the genetic intervals thought to contain the disease loci and
expedite their positional cloning, thereby providing a good chance of cloning the “diabetogenes” (25).

ACKNOWLEDGMENTS

We would like to thank Drs. Michael Boehnke, Francis Collins, Richard Bergman, Simeon Taylor, and Alan Permutt and two anonymous referees for significantly improving an earlier version of this paper.

APPENDIX: GLOSSARY

Admixture. A situation in which two populations with different allele frequencies at each of two separate loci (which are in equilibrium) are mixed. One consequence of admixture is that it can lead to a spurious association between marker loci and a trait for which the original population frequencies differ.

Allele. A gene or polymorphism at a specific locus that may exist in two or more states in a population.

Association study. An association between a particular genotype or allele (usually from a marker often closely linked to or within a gene) and a phenotype is said to exist when the particular genotype occurs more often than expected in individuals with the phenotype. To study association, one collects two groups of people, cases (i.e., those with the phenotype) and control subjects (i.e., those without the phenotype). In each group, the proportion of individuals with the interesting marker genotype is then measured and the frequencies are compared.

Assortative mating. Nonrandom noninbreeding matings (e.g., people of the same or similar heights marrying each other more often than people of different heights).

Asymptotic value. A statistical test that tends toward a mathematical expectation.

Branching process. A branching process is one type of stochastic process that is often used to study changes in a population over time. It can be used to analyze the behavior of a new mutation within a population after a given number of generations.

Complex disease. A trait that shows familial clustering but no clear Mendelian inheritance pattern. The lack of clear Mendelian inheritance is often due to the influence of multiple genetic and nongenetic factors.

Covariance. A statistical term that is a nonstandardized measure of the relationship between two traits. In particular, for traits X and Y covariance is \( \text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \), where \( \mu_X \) and \( \mu_Y \) are means for the two traits respectively and E is the mathematical expectation.

Epistasis. The nonadditive (i.e., interactive) influences of multiple genes controlling a phenotype.

Founder mutation. The mutation within a person (i.e., founder) who introduced the mutation into a particular population or pedigree.

Genotype. A pair of alleles at a specific locus constitutes a genotype. If the alleles are identical, the genotype is referred to as “homozygous” and if different as “heterozygous.”

Haplotype. The linear order of alleles at a number of loci received from one parent (or chromosome).

Heritability. Heritability refers to how much of the phenotype is determined by genes, and it is defined as the ratio of genetic variance, \( V_G \), to total phenotypic variance, \( V_P \), in its broad sense and as the ratio of additive genetic variance, \( V_A \), to total phenotypic variance in its narrow sense.

Informative polymorphism. This term refers to the existence of multiple alleles in a population for a marker such that for each mating the four grandparental chromosomes are clearly distinguished.

Heterogeneity (genetic). Refers to when any one of two or more genes can independently cause disease.

Likelihood. A measure of the joint probability of a number of observations given a particular model and parameter vector governing that model. The parameter in the likelihood equation may be estimated by the method of maximum likelihood, in which the likelihood function is maximized with respect to the parameter over the available data.

Linkage analysis. A statistical technique that is used to identify a specific chromosomal region containing a disease gene. While suited mainly for monogenic disorders, the basic idea behind linkage analysis is to trace and measure the co-segregation of disease in a family with marker loci by taking information from putative recombination events during meioses. The extent of linkage is measured by formulating a LOD score (see below), and the closer the markers are to the disease gene, the greater the extent of cosegregation and the larger the LOD score (generally). In standard linkage analysis, model parameters such as allele frequencies and penetrances have to be specified.

Linkage disequilibrium. Two closely linked loci are said to be in linkage equilibrium when the population frequency of a combination of alleles at these loci equals the product of individual population allele frequencies at the two loci. If they are not equal, the loci are said to be in linkage disequilibrium. Causes of linkage disequilibrium include introduction of a new mutation in a population, population stratification, and mixing of populations with differing allele frequencies for corresponding loci. Disequilibrium often exists over very short distances (say about a few hundred kilobases or less), and its existence is a good indication for the proximity of two loci or DNA markers.

Locus. Position of a gene, DNA segment, or polymorphism along a genome.

LOD score. Logarithm (base 10) of the likelihood ratio comparing assumptions of linkage and nonlinkage, i.e., under free recombination (\( \Theta = 0.5 \), where \( \Theta \) is the recombination fraction). Classically, a LOD score of 3 (1,000:1 odds for linkage) is the conventional cutoff point above which the significance of a linkage is declared (but see Land and Schork [64]).

Map distance (units of morgans). The map distance is the expected number of crossovers occurring on a single chromosomal strand between two loci. The length of a chromosome with an average of one crossover per meiosis is one morgan. A centimorgan (cM) is one one-hundredth of a morgan.

Marker. Usually refers to a polymorphic locus on a genome used for gene mapping purposes. Microsatellites (simple tandem-repeat structures) are the most abundant marker systems that are used in genetic linkage mapping.

Matrix. An array of numbers structured in dimensions.

Mixed model. A genetic model that assumes that a major gene, polygenes, and random environmental effects combine independently to produce the trait in question.

Mixture distribution. Refers to two or more probability functions (usually taken to be normal in genetic contexts) in
which the components are weighted by the relative frequencies with which individuals associated with each component occur.

**Multivariate normal.** Multivariate statistical analysis is concerned with data that consists of sets of measurements on a number of individuals. The multivariate normal distribution is uniquely determined by its mean vector (a column matrix) and covariance matrix. Many genetic analysis techniques assume each family's or pedigree's quantitative trait data follows a multivariate normal distribution.

**Path analysis.** A technique that allows one to represent in diagrammatic form linear structural models and hence devise predictions for the various covariances (the covariance structure) of variables under that model.

**Penetrance.** Probability of disease expression given a particular disease genotype.

**Positional cloning.** The method by which a gene is isolated on the basis of its location in the genome and not on the basis of any known function. Linkage analysis is primarily used to locate the gene to a specific chromosomal region. Finer mapping subsequently proceeds via linkage disequilibrium mapping and physical mapping where the region is cloned in overlapping contigs (contiguous segments). Special methods are then used to hunt for the disease gene among all other genes in this segment. Positional cloning is suited more toward mono-allelic disorders than polygenetic ones.

**Power.** The probability of rejecting a null hypothesis when the alternative hypothesis is true. Another way to express the power is the probability of true-positives that are picked up by the test.

**Phase and recombinant.** Consider two loci on the same chromosome with two alleles each, e.g., A and a for the first locus and B and b for the second. There are four possible haplotypes: AB, Ab, aB, and ab. If the individual has phase AB/Ab, that is, A and B have been inherited from one parental chromosome and a and b from the other, then the haplotypes AB and ab, which the individual passes to his/her offspring, are nonrecombinants, whereas Ab and aB are recombinants. The observed proportion of each haplotype will depend on the proximity between the two loci, with nonrecombinants being more frequently produced if the loci are close together (i.e., linked).

**Phenotype or trait.** The expression of a particular genotype or, in the context of complex traits, the effect of many, possibly unlinked, genotypes at separate loci in addition to the environment.

**Polymorphism.** A gene or other DNA segment is called a polymorphism when its most common allele has a population frequency of <95%.

**Proband.** An affected individual who has been detected independently of other members of the family and through whom the family can be ascertained.

**Recombination fraction (θ).** Refers to the probability that a gamete produced by a parent is a recombinant between two loci.

**Segregation analysis.** Statistical technique used to determine the mode of inheritance for a trait (e.g., dominant, recessive, due to single gene, multiple genes, etc.).

**Variance.** This term is a measure of the variation of the trait. If values are simply expressed as deviations from the population mean, then variance is the mean of the squared values. The genetic variance ($V_G$) is the variance in the phenotypic values attributable to genes, while the environmental variance is variation attributable to nongenetic or environmental factors ($V_E$). The total variance is the phenotypic variance and is the sum of these separate components ($V_T = V_G + V_E$). A component of genetic variance is the additive genetic variance ($V_A$), which reflects the variance of the average sum of individual gene effects about the population mean for one parent. The additive genetic variance is the chief cause of resemblance between relatives and the major determinant of observable genetic properties in the population.

**REFERENCES**


