Am J Pharmacogenomics 2001; 1 (3): 203-221
1175-2203/01/0003-0203/$22.00/0

Finding Genes Influencing Susceptibility to Complex Diseases in the Post-Genome Era

Bruce Rannala

Department of Medical Genetics, University of Alberta, Edmonton, Alberta, Canada

Contents

Abstract

During the last decade, hundreds of genes that harbor mutations causing simple Mendelian disorders have been identified using a combination of linkage analysis and positional cloning techniques. Traditional approaches to gene mapping have been largely unsuccessful in mapping genes influencing so-called 'complex' genetic diseases, however, because of low power and other factors. Complex genetic diseases do not display simple Mendelian patterns of inheritance, although genes do have an influence and close relatives of probands consequently have an increased risk. These disorders are thought to be due to the combined effects of variation at multiple interacting genes and the environment. Complex diseases have a significant impact on human health because of their high population incidence (unlike simple Mendelian disorders, which tend to be rare). New techniques are being developed aimed specifically at mapping genes conferring susceptibility to complex diseases. A project aimed at mapping genes influencing susceptibility to a complex disease may be undertaken in several stages: establishing a genetic basis for the disease in one or more populations; measuring the distribution of gene effects; studying statistical power using models; carrying out marker-based mapping studies using linkage or association. Quantitative genetic models can be used to estimate the heritability of a complex (polygenic) disease, as well as to predict the distribution of gene effects and to test whether one or more quantitative trait loci (QTLs) exist. Such models can be used to predict the power of different mapping approaches, but are often unrealistic and therefore provide only approximate predictions. Linkage analyses, association studies and family-based association tests are all hindered by low power and other specific problems. Association studies tend to be more powerful but can generate spurious associations due to population admixture. Alternative strategies for association mapping include the use of recent founder populations or unique isolated populations that are genetically homogeneous, and the use of unlinked markers (so-called genomic controls) to assign different regions of the genome of an admixed individual to particular source populations. Linkage disequilibrium observed in a sample of unrelated affected and normal individuals can also be used to fine-map a disease susceptibility locus in a candidate region. New Bayesian strategies make use of an annotated human genome sequence to further refine the position of a candidate disease susceptibility locus.

Complex diseases, according to Lander and Schork[1] are those that 'do not show perfect cosegregation with any single locus owing to such problems as incomplete penetrance, phenocopy, genetic heterogeneity, and polygenic inheritance.' Complex diseases are probably mainly polygenic, although many single locus Mendelian disorders also display one or more of these features. For example, for single locus disorders, incomplete penetrance can arise if environmental or developmental stochasticity leads to different outcomes for individuals with identical genotypes.[2] Incomplete penetrance is more complicated in the polygenic setting; the effect of no single gene may be sufficient to cause a disease phenotype and therefore the degree of penetrance for any particular locus is affected by the alleles segregating at other disease loci in a family, or a population. For a given disease susceptibility locus, individuals displaying the disease phenotype but lacking a disease mutation at that locus (phenocopies) may arise due to the effects of other genes, or the environment. Genetic heterogeneity can include both multiple disease genes (locus heterogeneity) and multiple mutations within a disease gene (allelic heterogeneity).

In this review, we focus on methods for mapping susceptibility loci for diseases arising from the combined effects of two or more genes, each with a potentially small (marginal) effect on the phenotype, as well as possible environmental effects. It is entirely possible that many traits are greatly affected by genes, but that each gene involved has a relatively small effect on its own. Thus, although the heritability of a trait may be high, the causative loci may be difficult to identify. A brief introduction is given to quantitative genetics, the branch of population genetics most useful for modeling such diseases. As well, the expected efficiency of different gene mapping strategies, using either linkage or association methods, is explored in the context of these models. Linkage mapping methods use information from recombination within families to identify markers cosegregating (and presumably genetically linked) with a disease locus. Association mapping methods examine marker alleles in affected and normal individuals to detect differences of allele frequencies between the two groups that may indicate either that a marker polymorphism is a cause of disease, a single nucleotide polymorphism (SNP) in a coding region for example, or that the marker is closely linked to a disease locus with which it is in population linkage disequilibrium (LD). If a marker allele is in LD with a disease susceptibility allele, it occurs more often on chromosomes bearing the susceptibility allele than would be expected at random (fig. 1).

Many problems need to be overcome before one can expect a reasonable chance of success in mapping genes underlying complex diseases. One complication has to do with the relationship between genotype and phenotype. In a majority of complex diseases, the disease phenotypes are highly variable and may involve measurements on many (possibly correlated) variables. Physicians may erroneously classify individuals as having different diseases when a common underlying disease displays high phenotypic variability. The opposite situation can also occur, with individuals affected by different disorders being placed in the same disease category. These problems reflect one of the more difficult aspects of studying complex disorders, understanding the relationship between phenotypes and genotypes. This relationship may be many-to-many, many-to-one, or one-to-many, unlike the case with simple Mendelian disorders.[3] In the absence of genetic heterogeneity, the relationship between phenotype and genotype for single locus Mendelian diseases is most often one-to-one. Even for simple Mendelian diseases, however, complications arise. Many examples exist of disorders that physicians have historically recognized as two or more syndromes, yet are later shown to be caused by a single disease gene.

The combinations of genes influencing a complex disease, and the magnitude of environmental effects (and therefore the heritability), may vary among families and populations. The

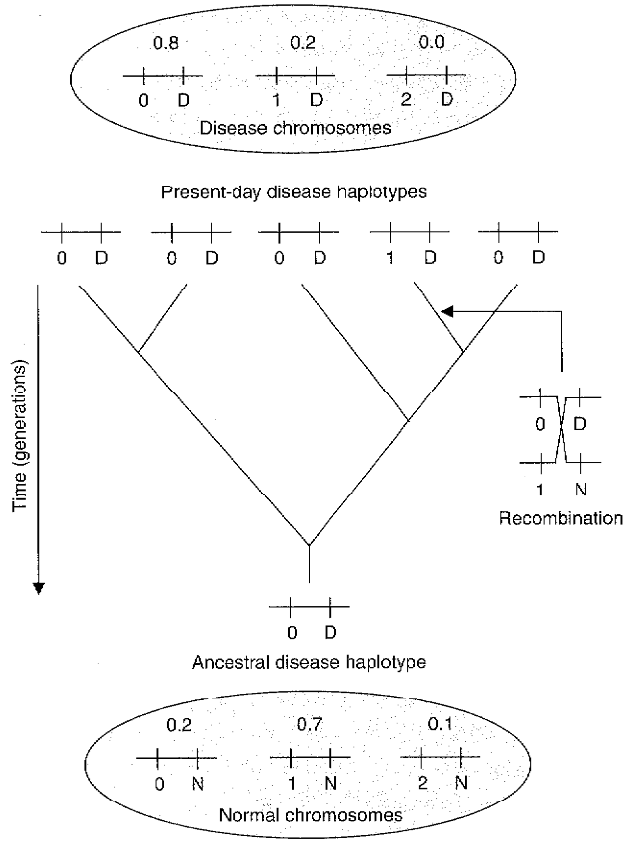

Fig. 1. Hypothetical population of 5 disease mutation-bearing chromosomes descended from a single ancestral chromosome illustrating the process by which marker allele frequencies are altered on disease-associated chromosomes versus normal chromosomes. The horizontal bars represent chromosomal haplotypes and the 2 vertical bars on each chromosome denote a disease locus (at right) and a marker locus (at left). N denotes a normal allele at the disease locus and D denotes a disease allele. Two alleles are present in the population at a linked marker locus, 0 and 1. The disease mutation arose on a chromosome bearing marker allele 0. The ellipse at the bottom of the figure symbolizes the population of normal chromosomes in which the disease mutation first arose. The population frequency is indicated above each haplotype. The ancestral haplotype (on which the disease mutation first arose) is shown at the bottom of the figure and above this a genealogy relating 5 descendent haplotypes. One recombination event occurred on this genealogy and is shown to the right of the figure. Time, in the past, increases from the top to the bottom of the genealogy. The ellipse at the top of the figure symbolizes the present-day population of disease chromosomes. The frequency (in the population of disease mutation-bearing chromosomes) is indicated above each haplotype.

effects of many genes may simply be too small to detect by conventional linkage or association techniques without unrealistically large sample sizes. As well, many of the populations from which individuals are sampled in studies of complex disorders are heterogeneous; admixture in such populations can lead to false associations of markers with a disease and may inflate estimates of heritability because of the presence of gametic phase disequilibrium.[4] Technical problems also arise in interpreting the results of whole-genome screens of thousands of markers in linkage and population association studies because the large numbers of markers used can result in many false associations unless very small type I error levels are used to determine significance, but this will reduce the power of the approach.[5] It may often be difficult to verify or reproduce associations observed in particular studies because differences in gene frequencies among populations can lead to high heritability (or strong association) in one population but not in other populations with different genetic backgrounds. Genetic stratification can affect linkage studies as well, causing the penetrance of a disease to vary among populations, or resulting in different susceptibility loci being mapped in different populations.

The ultimate usefulness of susceptibility loci, once identified, in genetic counseling or in the development of new therapies is also not clear. If genes have purely additive effects, it should be possible to accurately predict patient risk based on the genotypes observed at known susceptibility loci (and taking account of any environmental risk factors). However, if there are large epistatic effects between susceptibility genes, their relative effects may be highly dependent on the genetic context (depending on the overall genetic background) and therefore inherently less predictable. Similarly, if gene-environment interactions occur, individuals with identical genotypes may have different genetic risks if exposed to different environments.[6]

The above considerations suggest that strategies may need to be developed that are biased toward mapping susceptibility loci whose effects are additive. This would lead to the initial identification of genes most immediately useful in terms of predicting patient risk, or developing safe therapies of general use to patients carrying susceptibility genes. Genes whose effects are additive carry a fixed risk, allowing useful genetic risk predictions for individuals based only on their genotypes and the overall population risk. As well, potential drug therapies that modify the effects of an additive susceptibility locus should have predictable effects in reducing the risk of all patients with a given genotype. However, other, non-additive loci will still be of interest in treating patients in particular high-risk populations, and in understanding the overall nature of a disease, especially if followed up by gene targeting studies in model organisms.

The *APOE* ε4 susceptibility allele for sporadic Alzheimer disease presents an interesting example of a complex disease susceptibility locus with a nearly additive effect across genotypes. Corder et al.[7] found that the proportion of individuals unaffected by Alzheimer disease at age 75 years who carried the ε4/ε4 genotype was roughly 20%, whereas the proportion of ε3/ε4 individuals unaffected was roughly 40%, and the proportion of ε3/ε3 individuals unaffected was roughly 70%. Thus, each copy of the ε4 allele increases an individual's risk of developing Alzheimer disease by age 75 years by between 20% and 30% (see section 9).[8] The *APOE* ε4 allele has thus far been studied mainly in Europeans and may have less importance as a risk factor for Alzheimer disease in other ethnic groups.

The task of identifying susceptibility loci for complex diseases and elucidating their function is bound to be difficult, requiring a multifaceted approach using both mapping studies in humans and functional studies in transgenic model organisms. This research will be expensive and may require decades to complete. However, because such diseases are much more common than the rare genetic disorders that have been the subject of a majority of studies thus far, advances in this area have the potential to greatly influence humankind, alleviating the suffering of a large number of persons and extending the productive human lifespan. Time will tell whether this will be the greatest contribution of human genetic studies.

Given their enormous potential for diagnosing and treating human disease, few would argue that genes affecting complex diseases are not worth pursuing, despite some discouraging obstacles. However, recent reviews tend to be either very pessimistic[9] or very optimistic.[3] Here, the aim will be to present an overview of current strategies for mapping complex disease that falls somewhere between optimism and pessimism, discussing the most significant difficulties that arise in studies aimed at mapping genes influencing complex traits and outlining some existing strategies aimed at overcoming these problems. The ultimate usefulness of many of these strategies has yet to be proven, as genetic studies of complex diseases are still in their infancy.

The difficulty and expense of undertaking dense marker screening and mutation detection studies to identify genes influencing susceptibility to common diseases requires that a prospective study be approached cautiously; evaluating the overall feasibility at each stage. A possible progression of strategies is as follows: establish that a disease is significantly influenced by genes and is not purely due to environment; examine the distribution of gene effects and assess whether evidence exists for disease susceptibility loci of major effect; examine the potential power and/or feasibility of different gene mapping approaches (i.e. linkage analysis, association studies, etc.) to detect susceptibility loci; undertake genotyping studies of single nucleotide polymorphisms, or microsatellite markers, aimed at identifying disease susceptibility loci (or candidate regions for susceptibility loci). All of these objectives, apart from the last, can be accomplished prior to a molecular genetic analysis; if the results are predominantly negative this could be a basis for precluding a marker-based gene mapping study. Thus, although the aim of this paper is to describe marker-based methods for mapping genes influencing common diseases, a review will first be presented of methods for establishing whether a complex disease is influenced by genetic variation segregating in a population and whether a genotyping study is likely be successful in uncovering susceptibility loci.

1. Genetic Models of Complex Diseases

To establish the role of genes in causing a complex disease, and the potential power of a particular mapping strategy to find disease susceptibility genes, mathematical models are needed that allow one to predict phenotype from genotype. Simple models of Mendelian segregation typically cannot account for the pattern of recurrence of a complex disease in families; this is expected if such disorders are polygenic and affected by environment. The relationship between phenotype and genotype is complicated in such cases, yet some very simple models have been developed over the last century that can be quite useful and provide a starting point for addressing this relationship. In this section, we describe these models, which fall into the realm of 'quantitative genetics', and their application. For a more comprehensive introduction to quantitative genetics, see the recent books by Falconer and Mackay[10] and Lynch and Walsh.[11]

1.1 Historical Overview

Mathematical models of complex traits were developed shortly after the rediscovery of Mendel's work. The models were needed to reconcile the observations of biometricians such as Johannsen that the variation of many traits, such as weight and height, is not discrete like the pea traits studied by Mendel, but instead appears continuous, and is often normally distributed in populations. Initially, this was taken as evidence against a discrete gene model of inheritance and alternative models of blending inheritance were developed to explain the observations. Darwin[12] had assumed that continuous characters exhibited blending inheritance and Galton[13] and Pearson[14,15] further developed this theory. Fisher[16] showed that a blending inheritance hypothesis was inadequate to explain existing variation in populations because such a process would eliminate variation too quickly, halving the genetic variation with each generation of random mating.

Fisher[17] proposed a model for the genetics of continuous characters in which a trait is governed by many genes and influenced by environment. This theory provided a mathematical formulation of earlier ideas, arising from experimental studies by Mendel and others, that continuous variation could result from the effects of multiple independently segregating genes.[11] A polygenic discrete gene model, with each gene having an independent additive effect (of similar magnitude) on the phenotype, results in a normal distribution of the phenotype, agreeing with the observations of the biometricians.

1.2 The Basic Model

The area of population genetics that applies Fisher's theory to study continuous traits has come to be known as 'quantitative' genetics, reflecting the fact that the traits under study are typically measured rather than placed into discrete categories as with simple Mendelian traits. The canonical model employed in quantitative genetics treats the phenotype of an individual as determined by an equation of the form

$$P=\sum_i \sum_j x_{ij} + D + I + \varepsilon, \qquad \text{(Eq.1)}$$

where x_{ij} is the additive contribution of allele j at locus i to the phenotype (P), D summarizes the effects of all non-additive interactions between alleles at the same locus (dominance effects), I summarizes the effects of all the non-additive interactions between alleles at different loci (epistatic effects), and ε is the effect of environment, typically assumed to be a random variable from a symmetrical distribution with a mean of zero. Figure 2 illustrates the population frequency distribution (obtained by simulation) of phenotypes (in a sample of 1000 individuals) for a complex trait influenced by either 4 loci or a single locus, with each locus having 2 alleles in equal frequency, and with each allele either adding or subtracting 1 unit from the phenotype. Environmental effects were assumed to follow a normal distribution, with a mean of zero and a variance of either 0.25 (for the case of a single gene affecting the trait) or 1.0 (for the case of 4 genes affecting the trait). Even with only 4 genes affecting the trait, the frequency distribution in the population approaches a normal distribution.

1.3 Disease Threshold Models

The theory outlined above deals with continuous, normally distributed traits. Most complex diseases do not fit this description, although the underlying phenotype is often continuous at some level. On the one hand, physicians tend to recognize individuals as affected by a disease displaying a continuous range of phenotypes in a population if their phenotype falls outside of what is considered the normal range (for example, obesity). On the other hand, a polygenic disease might appear discrete, but this is because it is only manifest when the underlying phenotypic variables exceed some biological threshold (cleft palate is a possible example). Threshold models can be usefully applied in both cases described above and were used early on to model discrete variants that did not display a simple pattern of Mendelian segregation by Pearson,[15] Wright,[18] Dempster and Lerner,[19] and others.[20] Edwards,[21] Mendell and Elston,[22] and others applied these models in estimating familial disease risk in humans. The models relate the discrete incidence of a disease to an underlying continuous model of gene effects by assuming that a threshold value exists for the trait; if an individual's value exceeds this

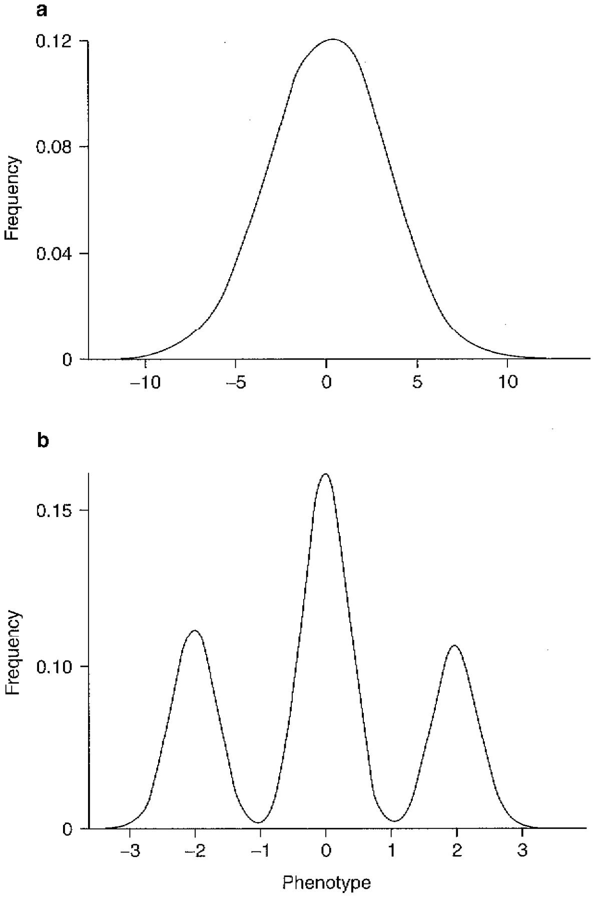

Fig. 2. The population frequency distribution of a phenotype in each of 2 simulated diploid populations, each comprising 10 000 individuals. **(a)** A simulated population in which it is assumed that the trait is determined by 4 biallelic loci, with each allele having an additive (either +1 or −1) effect, and an environmental influence that is normally distributed with mean 0 and variance 1. The phenotype distribution is unimodal, and approaches a normal distribution, despite the relatively small number of loci influencing the trait. In this case, the trait appears complex. **(b)** A simulated population in which it is assumed the trait is determined by a single biallelic locus, with each allele having an additive (either +1 or −1) effect, and an environmental influence that is normally distributed with mean 0 and variance 0.25. In this case, the distribution is tri-modal; the left and right modes represent the homozygous (−/− and +/+) phenotypes (deviations about −2 and +2 are due to environment) and the middle mode represents the heterozygous (−/+) phenotype. In this case, the trait appears Mendelian.

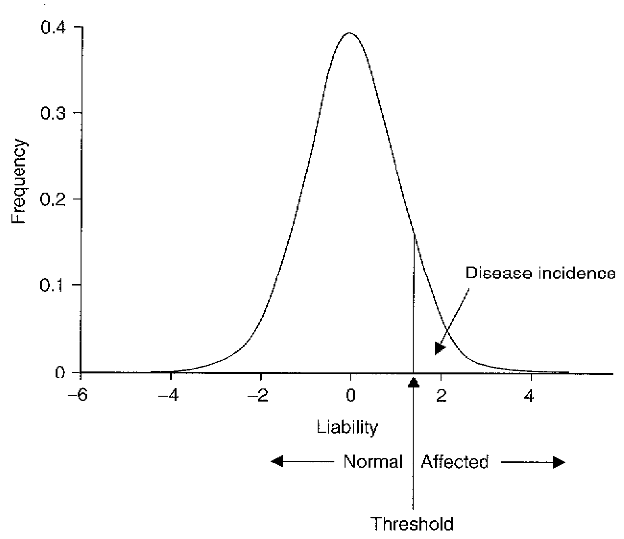

Fig. 3. A hypothetical threshold trait. The curve represents the population frequency distribution of liabilities (underlying phenotypes), the vertical line at the right is the threshold value for the trait. Individuals with a liability value to the right of the threshold are affected, those with a liability value to the left are unaffected. The population disease incidence is the area under the curve to the right of the threshold value.

threshold they display the disease phenotype, otherwise they do not. This concept is illustrated in figure 3.

In disease threshold models, the horizontal axis, which measures the trait value, is referred to as the liability. The models are therefore often referred to as liability models. The area under the curve for liability values exceeding the threshold is the expected incidence of the disease in a population (fig. 3). Although threshold models have played an important role in human genetics, many disorders have a broad range of phenotypes for which any designation of 'diseased' versus 'normal' is somewhat arbitrary. This allows scope for the power of mapping studies to be increased by altering the categories used to assign individuals to the disease class. Classical results relate the strength of artificial selection to the change in population allele frequency following an episode of selection.[11] Based on these results, a number of researchers have argued that allelic associations are more easily detected by choosing more extreme phenotypes (i.e. raising the threshold value),[23] possibly by sampling from both the upper and lower tails of the phenotype distribution.[24] This has been advocated as a general strategy for mapping genes influencing complex disorders (but see Allison et al.[25]).

There is general agreement that phenotypes must be more carefully identified and subcategorized for studies of complex diseases. One novel approach is to compare the severity of the phenotypes of affected individuals with different genotypes at a marker locus; if a difference is detected this can indicate a role for the locus (or a nearby linked locus) in disease.[26] Since most threshold models are based on univariate phenotypes and realistic disease categories will usually be multivariate, more studies are needed that consider the effects of ascertainment to the disease class on the power to detect allelic associations. In the case of complex diseases displaying a continuous range of phenotypes, measurements on disease phenotypes may be analyzed directly in gene mapping studies. Methods have been developed for pedigree-based linkage mapping of a quantitative trait locus (QTL) using a variance components approach with either a single linked genetic marker[27] or multiple genetic markers[28] and analyzing measurements of a continuously varying trait. The term QTL is often used to describe a major locus that accounts for a large proportion of the population variance of a quantitative trait phenotype.

2. Disease Heritability

Early applications of quantitative genetics to human disease focused mainly on measuring the degree of genetic determination for particular disorders. This work relies heavily on the variance component analysis approach developed by Fisher.[17] A review aimed at statistically-oriented readers can be found in Hopper.[29] The basic aim of variance component methods is to partition the population variation of a phenotype (V_P) into components arising from different sources. The main dichotomy of interest to geneticists separates variance due to genetic differences among individuals (V_G) from variance due to differences of environment (V_E). The genetic component of variance can be further subdivided into additive (V_A), dominance (V_D) and epistatic (V_I) genetic variance as illustrated in the following equation[4]

$$V_P = V_A + V_D + V_I + V_E. \qquad \text{(Eq. 2)}$$

These variance components correspond to the various genetic and environmental contributions presented in equation 1 (section 1.2). The 'additive' genetic component refers to the variance arising from the terms involving x_{ij} in the sum of equation 1. The magnitude of each component of genetic variance depends on population allele frequencies, as well as gene effects. A gene with a large effect, for example, may contribute little to the genetic variance if it is in low frequency. Components of genetic variance have been used to define measures of the heritability of a trait. The broad sense heritability is defined as $H^2 = V_G/V_P$ and the narrow sense heritability as $h^2 = V_A/V_P$. H^2 has been of most interest to human geneticists, while h^2 has been of interest to animal breeders.

2.1 Estimating the Heritability of a Complex Disease

Before attempting to identify genes for a complex disorder, it is prudent to first determine whether there is a significant genetic component for the disease in a study population. Many techniques have been developed for estimating the heritability of complex genetic diseases. One common approach uses studies of pairs of close relatives, or of monozygotic (MZ) or dizygotic (DZ) twins. The usual approach in twin studies is to compare the correlation of phenotypes between MZ versus DZ twins.[2] Because MZ twins share all their genes, while DZ twins share only half their genes, the difference in the correlation of trait phenotypes between MZ versus DZ twins can be used to estimate trait heritability. The broad sense heritability is estimated as twice the difference between the correlation coefficient of the quantitative phenotype in MZ versus DZ twins. Twin studies potentially offer experimenters greater control over environmental and genetic components of phenotypic variance and are therefore very promising for studies of complex disease.[30] Another approach to estimating heritability evaluates the patterns of transmission of a quantitative trait (or disease) phenotype on extended pedigrees.[31]

2.2 Disease Heritability and Population Genetic Structure

Measures of heritability are influenced by population genetic structure because of the dependence of heritability measures on population allele frequencies at the relevant loci. For example, in the case of a single genetic locus, with two alleles A_1 and A_2, with population frequencies p and $1 - p$, respectively, and with A_1A_1 having phenotype a, A_1A_2 having phenotype d, and A_2A_2 having phenotype $- a$, H^2 is

$$H^2 = \frac{V_A + V_D}{V_A + V_D + V_E}$$

and[4]

$$V_A = 2p\,(1-p)[a + d(1-2p)]^2,$$
$$V_D = (2p(1-p)d)^2.$$

The effect on H^2 of changing the population frequency, p, of allele A_1, determined using the above equations, is illustrated in figure 4. If alleles with additive effects influencing a disorder have either a very low or a very high frequency in a population, the disease will have low heritability in that population. As a consequence, a particular genetic disease may have high heritability in one population and low heritability in another. A lack of reproducibility (in additional populations) of the result of a study indicating high heritability of a trait cannot be taken as evidence against a genetic component to the disorder; the safest conclusion is that either there is little segregating variation for the disease loci in these additional populations, or there is an increased influence of the environment.

2.3 Reliability of Disease Heritability Estimates

Several kinds of interactions can inflate the heritability of a trait even though limited additive and epistatic genetic variance exists. These include gene-environment interactions,[6] which can cause heritability estimates to depend on environmental effects (that may vary among populations), and gametic phase disequi-

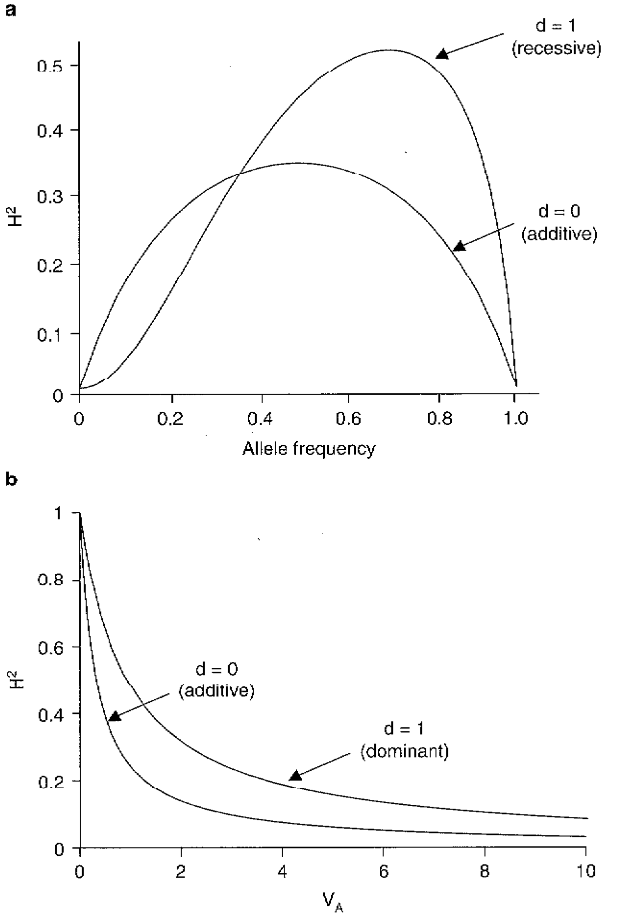

Fig. 4. The effect that changing allele frequencies, and environmental variances among populations will have on the broad sense heritability of a trait (H^2), determined by a single locus with either additive, recessive, or dominant effects among alleles. It is assumed that the locus is biallelic and the relative contribution of each of the 3 distinct phenotypes (homozygote -/-, heterozygote +/- and homozygote +/+) to the genotype are −1, d and +1, respectively, where d is the phenotypic value of the heterozygote. **(a)** H^2 as a function of allele frequency for either recessive, or additive, effects among alleles [assuming the environmental variance (V_E) is 1.0]. Populations with intermediate allele frequencies will display the highest heritabilities for the trait. **(b)** H^2 as a function of V_E for either recessive, or additive, effects among alleles (assuming the allele frequency is 0.5). Populations with greater environmental variance will show reduced heritabilities for the trait.

librium, which can inflate genetic variance and thus heritability estimates.[4] Variance due to gametic phase disequilibrium (often due to recent admixture) may be particularly problematic when heritability estimates are derived from pedigrees that may include admixture. These factors together suggest that heritability estimates should be taken as only crude indicators of the role of genes in producing disease in any given population. Twin studies have been carried out to estimate heritability for many complex traits and estimated heritabilities are often exceptionally large. For example, Commuzzie and Allison[32] estimated heritabilities for several components of obesity to range from 40% to 70%. In general, most factors bias estimates in favor of larger heritabilities and so these results should be interpreted as placing an upper bound on the heritability of a trait.

3. Measuring the Distribution of Gene Effects

If heritable variation exists for a disease trait in a population, one can use the phenotypes observed either on extended pedigrees, among parents and offspring, or among sib-pairs, to assess the evidence for one or more loci of major effect (QTLs). The method of complex segregation analysis (CSA) has been used to carry out such studies in humans.[33,34] Existing CSA methods typically use likelihood ratio tests to evaluate hierarchical models of either no genetic effects, one major gene with no background polygenes, no major genes but only background polygenes, or one major gene plus background polygenes.[11] Recently, these approaches have been extended in various directions. In particular, the calculations can now be carried out on large pedigrees using Markov chain Monte Carlo methods.[35] CSA methods should be used with caution as they can be very sensitive to model assumptions. In particular, the methods assume that the distribution of phenotypes of individuals having a given genotype is normal. If this assumption is violated, the tests may falsely indicate a major disease locus.[36,37] Another problem is that, if genotype-environment interaction is present, the power of the methods to detect a major gene may be very low.[38,39]Another approach to determining whether a disease is caused primarily by a single major locus, or is instead polygenic, compares the disease recurrence risk ratio λ for different classes of relatives (siblings, cousins, etc.). The risk ratio for a type R relative of an affected individual is the probability that the relative is affected divided by the probability that a randomly chosen individual from the population is affected (i.e. the population disease prevalence). Under a range of models with either additive, or epistatic, effects among disease alleles and loci, the value of λ decreases more rapidly with an increasing distance of relationship if multiple disease loci are involved.[40] Based on this, Risch[40] argued that a single genetic locus was not compatible with existing data on the incidence of schizophrenia among relatives of various degree. Approaches to detecting polygenic diseases based on risk ratios should be interpreted with caution because gene-environment interactions can produce similar patterns (of decreasing λ with increasing distance of relatedness), as would be predicted in the presence of polygenes even when a single major disease locus exists.[6]

4. Predicting Mapping Power Using Quantitative Genetic Models

Several authors[40-42] have considered the power of either linkage, or association, methods for mapping complex disease loci under very specific quantitative genetic models. These include either models in which disease alleles have purely additive effects within and among loci,[42,43] or multiplicative models of epistasis in which the penetrance of a phenotype is the product of the penetrance factor of each allele at each susceptibility locus.[43] Most of these models are special cases of a general model described by James.[44] Although the models provide a rough guide to the performance of different methods, they are obviously quite artificial and the results may not generalize. In this review, we consider the power of different gene mapping methods only in the context of a few simple models. Although a range of models have been developed over the last century to describe complex traits, none are entirely realistic and most are quite unrealistic. Many genes may display complex higher-order epistatic interactions, be influenced by gametic phase disequilibrium due to population substructure, or be subject to gene-environment interaction.

More effort needs to be devoted to studying the robustness of current methods for modeling quantitative traits and exploring the possibility of developing new models that are more robust. Developing very intricate models accounting for all the possible interactions is clearly hopeless, except for traits affected by a handful of loci. So-called 'oligogenic' models endeavor to study more complex models by limiting the number of major genes.[45] An important question for human geneticists is the form of the distribution of gene effects across disease susceptibility loci (fig. 5). If most complex diseases are influenced by one or more major genes in most populations there is more hope that these will be identified and effective treatments developed based on these findings.

5. Methods for Mapping Genes Influencing Complex Diseases

Two distinct approaches, linkage mapping and association analysis, have been used to map disease mutations. Linkage mapping uses inferred recombination on pedigrees made up of affected and

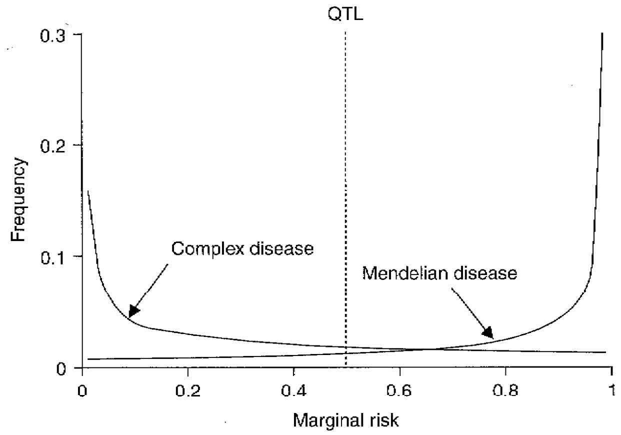

Fig. 5. Hypothetical frequency distribution (in the population of affected individuals) of gene effects across the loci that influence a polygenic (complex) disease, or a Mendelian disease, as measured by the marginal disease risk due to each locus (i.e. the probability that an individual has the disease given that they possess a susceptibility allele at the locus; considering only heterozygotes). The dashed line indicates one possible (arbitrary) criterion for identifying a locus as a quantitative trait locus (QTL), if the marginal risk due to alleles at the locus exceeds 0.5. For a complex disease, the frequencies (in the population of affected individuals) of disease-associated alleles at loci with large effects (QTLs) will be low (right of figure), while the frequency of disease-associated alleles at loci with small effects will be high (left of figure). For a Mendelian disease, the frequencies (in the population of affected individuals) of disease-associated alleles at QTLs will be high (right of figure), while the frequency of disease-associated alleles at loci with small effects will be low (left of figure).

normal individuals to identify markers that are closely linked to a disease locus. Although linkage mapping methods have been used successfully to map genes associated with complex diseases, the *APOE* locus contributing to late-onset Alzheimer disease being a notable example,[7] the power of linkage methods to map genes for diseases with complex inheritance (and low gene-specific penetrance) is often low. When carrying out linkage analysis using affected sib pairs, for example, the sample sizes needed for 80% power (i.e. to reject the null hypothesis of a linkage when it is false) are generally greater than realistically possible.[5]

Association methods for mapping disease susceptibility loci typically compare allele, or haplotype, frequencies between samples of affected and unaffected individuals. If an allele (or haplotype) has a higher frequency in affected versus unaffected individuals this may indicate that it contributes to the disease in the population, or is in linkage disequilibrium with a disease susceptibility locus. Association studies are generally more powerful than linkage methods when locus-specific disease penetrance is low, but can be prone to false-positives when population stratification exists, or controls and affected individuals are improperly matched. Linkage disequilibrium (LD) can also be useful for high-resolution mapping of a disease susceptibility locus once a candidate region has been identified. LD mapping methods use the expected excess in frequency among affected individuals of the marker haplotypes on which disease mutations first arose to fine-map the location of one or more disease susceptibility genes.

Here, we give a brief overview of current linkage and association mapping strategies and their relevance for mapping genes affecting complex genetic diseases. As well, we consider how the availability of a human genome sequence is impacting genetic studies of complex diseases using these techniques.

6. Linkage Mapping and Complex Diseases

Early linkage methods were based on direct counts of recombination events; the number of observed recombination events is divided by the number of informative meioses to directly estimate the recombination fraction between a marker locus and a disease mutation. For small distances [less than 10 centiMorgan (cM)] the fraction of recombinants, θ, provides a good estimate of the map distance, x, while for larger distances the relationship is more complex. A model first proposed by Haldane,[46] which assumes no hotspots or interference, specifies the relationship between the absolute value of the map distance and the expected recombination fraction. The Haldane model takes account of multiple cross-overs and is based on a Poisson process model of recombination; several more complex models, allowing for interference and other sources of non-independence among recombination events, have been proposed.[47]

For humans, the sex averaged relationship between physical distance and map distance is roughly 1 cM = 1 Megabase (Mb).[48] Taking advantage of this relationship, recombination fractions estimated by linkage analysis can be used to localize a disease mutation to a physical region of a chromosome. Direct approaches for estimating recombination fractions do not make full use of the available data. Because individuals are genotyped, and not haplotyped, phase is unknown and must be inferred from the transmission patterns of markers on pedigrees. In some cases, this can be done unambiguously and a direct approach is fully informative. In most cases, however, the phase of many chromosomes in the pedigree cannot be unambiguously determined, although a restricted number of possibilities may exist. Mathematical models can weight these different possibilities according to their probabilities and allow information to be extracted that would be unavailable to the direct method. In addition, direct approaches are prone to biases which are avoided by the use of statistical models.[49]

6.1 Pros and Cons of Linkage Mapping in Studies of Complex Diseases

Many statistical techniques have been developed over the last several decades for estimating the recombination fraction between a marker and disease locus (enabling the genetic and physical distances between them to be predicted) using probability models of allelic transmission and recombination.[49] These methods are typically either based on the method of maximum likelihood,[50,51] or use a Bayesian strategy.[52] Originally, the methods allowed only one marker to be analyzed, but recent theoretical developments, most notably peeling algorithms developed in the 1970s,[33] and computer programs allow multipoint linkage analysis to be carried out to simultaneously estimate recombination fractions between multiple linked markers.[49]

Linkage mapping methods have proved highly effective for mapping mutations in genes that cause simple Mendelian disorders. A major strength of the methods, when applied to rare genetic disorders, is that they are insensitive to allelic and locus heterogeneity. This is because multiple disease alleles, or genes, will rarely occur within families when the population incidence of a disease is low.[3] Disadvantages of linkage mapping include greatly reduced power to detect disease alleles that have low penetrance and a limited resolution. Even when large extended families are available, only a few hundred informative meiotic events can be observed, limiting the resolution of linkage mapping to 1 cM (roughly 1% recombination) or less.[53] Other approaches based on allele sharing (or genomic identity-by-descent) between relatives[54,55] have similar limitations, although larger sample sizes can often be obtained when only small clusters of relatives are needed (affected sib pairs for example).

6.2 Power and Significance in Linkage Analysis of Complex Diseases

Many studies have examined the power of linkage analysis, and the sample sizes needed to map genes influencing complex diseases. Most studies have considered affected-relative-pair designs, rather than extended pedigrees[56,57] and have focused on the sample sizes required either to detect an association with a marker conferring a particular genotype relative risk[57] or to reduce the size of the candidate region to less than 1 cM, allowing positional cloning.[56] With the availability of a human genome sequence, the size of a candidate region is now less critical. Other studies have focused on the significance levels that should be used in whole genome screens for linkage with a disease. This is particularly important for complex disorders because the sample sizes needed to detect significant linkage may be very large. For example, Lander and Kruglyak[57] offered several extended definitions of linkage for use in whole genome linkage studies that correct for marker number in detecting linkage to complex disease loci. Linkage is considered 'suggestive' if a false-positive would be expected to occur once, on average, in a whole-genome scan, it is considered 'significant' if a false-positive result would be expected for 5% of whole-genome scans, and it is considered 'highly significant' if a false-positive would be expected for 0.1% of whole-genome scans. The specific logarithm of the likelihood of odds (LOD) scores that correspond to these false-positive rates depend on the study design, but for a range of allele sharing methods in humans, suggestive linkage was indicated by an average LOD score of about 2.25 and significant linkage by an average LOD score of about 3.65. These values do not differ greatly from the classical criterion for significance in single locus linkage analysis of a LOD score of 3.[58]

6.3 Choice of Markers for Linkage Analysis of Complex Diseases

Other questions arise in linkage studies of complex diseases relating to the number of marker loci that should be included in a whole genome screen, and whether highly polymorphic microsatellite markers, which are less common throughout the genome than are the less polymorphic SNPs, should be preferred because they carry more information. Carrying out simulation studies of the inheritance information extracted using markers with varying degrees of polymorphism, and at varying densities, Kruglyak[59] concluded that SNPs could be as informative as microsatellite markers, given a sufficiently dense map. He also found that there was a limit (determined by the sizes of families) to the information increase (for linkage methods) obtained by increasing the density of markers used in a whole genome screen and concluded that marker densities of one per cM or less were sufficient for an initial screen for linkage.

7. Association Studies and Complex Diseases

Another strategy for identifying linked markers or disease susceptibility loci compares marker allele frequencies between unrelated affected and control individuals. The basic idea is that causal polymorphisms, or alleles at markers very closely linked to a disease locus, will occur in higher frequency in affected versus unaffected individuals. The simplest approach is to use a χ^2 test to compare the allele frequencies of markers between groups of unrelated affected and normal individuals. A significant result indicates that the underlying allele frequencies are different, for a given marker, in the two groups. If each sample is an independent random sample (apart from the stratification by disease state) and a single marker is analyzed, then significance indicates that either the polymorphism is a disease locus, or it is linked to a disease locus. Recent studies suggest that association methods can be much more powerful than linkage methods for identifying disease mutations with low penetrance.[5] One reason is that unrelated individuals share many fewer markers than relatives and shared alleles among affected, versus unaffected, individuals are therefore more informative.

7.1 False Associations, Multiple Tests and Population Admixture

At least two problems can confound association studies of disease susceptibility loci. First, as mentioned earlier, many marker loci must be analyzed and therefore many tests are performed and a correction for multiple tests must therefore be used.[60] Second, genetic drift and variable demographic histories for different subpopulations can result in variable frequencies of both a disease and a neutral marker among subpopulations. If, by chance, a marker allele and a disease both occur in high frequency in one subpopulation and in low frequency in another, then an association study of individuals from a population that is an admixture of the two can result in a significant association between the marker and the disease although they are in fact unassociated within each subpopulation. In this case, both the marker and the disease are simply indicators of population affiliation. This concept is illustrated in figure 6. Most associations between markers and disease identified to date have not been reproduced in subsequent studies, indicating either a high rate of false-positive results for such tests, or variable frequency of susceptibility genes and/or linkage disequilibrium among populations.

7.2 Family-Based Association Tests

A number of alternative methods have been developed for identifying disease-allele associations in the presence of population admixture. The transmission-disequilibrium test (TDT),[61] and the many recent variants of this method,[49] use the alleles transmitted from parents to affected offspring as the cases and the alleles that are not transmitted as the controls. The assumption is that both alleles of a parent are from the same subpopulation and are therefore appropriately matched controls for each comparison. This is true only if the parents are not themselves admixed. An assumption of the methods is therefore that admixture occurs only between parents and families and is not present within the genomes of individual parents. Although these methods show considerable promise, and are particularly attractive because they are non-parametric (not depending on the details of a particular genetic model), the power has turned out to be disappointingly low for the sample sizes used in many studies, especially for alleles with dominant effects.[62]

7.3 Association Mapping in Homogeneous Populations

Association studies that do not require relatives have at least two major advantages over methods that use relatives as controls. First, they enable much larger sample sizes to be obtained because DNA is needed only from affected individuals. Second, fewer ethical issues arise (samples and consent forms are needed only from the individuals initially recruited to a study). Because of the larger sample sizes that are possible, population association meth-

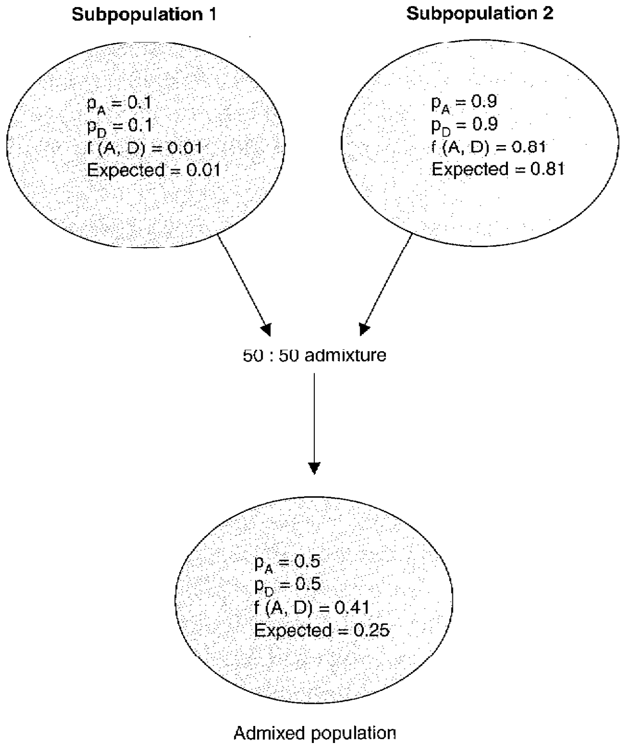

Fig. 6. Illustration of the way in which population admixture can lead to a spurious association between a marker and a disease. The two ellipses at the top of the figure represent hypothetical subpopulations (1 and 2). The marker allele frequency, p_A, and the disease frequency, p_D, in each subpopulation is indicated, as well as the expected frequency of affected individuals who carry the marker allele (based on the marginal frequency of each and assuming they are independent) and the actual frequency, denoted f(A,D). For the two source subpopulation, it is assumed that no linkage exists between the marker and the disease (which could be purely due to environment). The ellipse at the bottom of the figure represents a population generated by 50 : 50 admixture of the two subpopulations. For simplicity, it is assumed that it is the first generation of admixture and no matings have yet occurred within the admixed population. The marker allele frequency, and disease frequency in the admixed population are shown, as well as the actual frequency of affected individuals carrying the marker, and the expected frequency, assuming no linkage (and ignoring admixture). The actual frequency is much higher than expected and (without taking account of admixture) would lead to a false conclusion that the allele is associated with the disease.

ods can be expected to have increased power over TDT-type methods that use relatives as controls. As a consequence, there has been much recent interest in developing techniques for carrying out association studies that do not require the use of relatives as controls.

One way in which association studies can be legitimately done using unrelated cases and controls is to choose individuals from isolated, or founder, populations. These individuals will have a common demographic history reducing the chance of spurious associations due to population admixture and reduced genetic heterogeneity.[63] In addition, recent founder populations may display more LD between a disease locus and linked markers[64] although recent evidence suggests that the difference in background LD between isolated versus more heterogeneous populations may be modest.[65] The population of Finland is a good example of a founder population, having undergone a founding event roughly 2000 years ago, followed by a rapid expansion with limited immigration and increased endogamy within subpopulations.[66]

7.4 Association Mapping in Founder Populations versus Ancient Isolates

At least two strategies exist for gene mapping studies in homogeneous populations: (1) choose an isolated founder population that may be expected to display higher levels of LD and reduced genetic heterogeneity for a complex disease (possible examples include Finland and Iceland); (2) choose an ancient isolated population that has remained relatively small (and constant) in size for many generations, relying on occasional population bottlenecks, and genetic drift, to reduce genetic heterogeneity for a complex disease and increase population LD.[64] It is still too early to predict which (if either) strategy will be effective. Currently, several commercial and academic efforts are underway to map disease genes in homogenous populations such as those of Iceland[67] and Norfolk Island[68] taking advantage of these potential effects.

7.5 Association Studies in Heterogeneous Populations Using Genomic Controls

Another way in which such studies can be done, in large heterogeneous populations, avoiding false associations due to admixture, is to use a set of unlinked neutral markers to identify individuals with different levels of admixture.[69-71] The basic principle of these methods is to stratify heterogeneous populations into homogeneous subgroups based on the variation observed at unlinked neutral markers. Methods for carrying out association studies in heterogeneous, or admixed, populations using genomic controls are only now being developed and (at the time of writing) have not been widely used in actual gene mapping studies; the methods hold considerable promise with recent simulation studies suggesting that they may be more powerful than a TDT test when population stratification is not too extreme.[72]

7.6 High Resolution Mapping of Disease Loci by Linkage Disequilibrium

LD mapping methods use samples of unrelated affected (and normal) individuals to carry out high resolution mapping.[73-77] The basic idea is that the population genealogy underlying a sample of chromosomes from unrelated affected individuals is much larger than typical extended pedigrees, thus allowing many more opportunities for recombination to occur and providing a higher-resolution map. Thus far, LD mapping methods have been most effective when applied to studies of isolated founder populations such as the Finnish.[66] Most studies have used LD mapping to narrow the candidate region for a disease gene in a mapping project immediately prior to positional cloning phase.[77] With the availability of a human genome sequence,[78,79] positional cloning becomes unnecessary because genes in a candidate region can be identified and directly sequenced and analyzed for mutations. In the past, with severe Mendelian disorders this has usually meant looking for nonsense, or other, mutations present in affected but not unaffected individuals. With complex diseases, interest will instead focus on missense (and other more minor) mutations, which are present at increased frequency in affected versus unaffected individuals. Ideally, the effects of any suspected disease susceptibility mutations detected in this way will then be studied *in vitro*, or in model organisms, by gene targeting or other techniques.

LD mapping techniques will play a new role in the post-genome era; they will no longer be needed for positional cloning, but they will still remain useful for assigning probabilities that particular genes in a candidate region (identified from a human genome sequence) are disease susceptibility loci. This can increase the rate at which susceptibility loci are identified by greatly reducing the number of genes that will need to be sequenced for polymorphisms in the candidate region. Recently, studies of population LD have been advocated for direct use in whole genome association studies to map genes influencing complex diseases.[80] The effectiveness of population LD for high-resolution mapping of disease genes in a candidate region is now well-proven.[66] However, it is not yet clear how useful LD will be in whole-genome marker studies aimed at finding susceptibility genes for complex diseases. It is entirely plausible that LD may not extend over large enough regions to be of use in identifying disease loci in genome-wide scans (using present marker map densities) but will nonetheless be extensive enough to aid in high-resolution mapping, even in heterogeneous populations, once a candidate region has been identified using other approaches (such as linkage analysis). Several authors have recently proposed hybrid methods for mapping genes using linkage and LD information jointly.[81,82] These methods are too new to allow any conclusions here regarding their usefulness.

8. Exploiting a Human Genome Sequence in Studies of Complex Diseases

Two strategies have been proposed for mapping genes influencing complex diseases by population association. Both aim to take advantage of an annotated human genome sequence now available.[78,79] One strategy uses a dense set of single nucleotide polymorphisms (SNPs) within coding, regulatory and other functionally significant regions in which mutations influencing complex diseases are most likely to occur.[5] The basic idea is that by comparing frequencies at these loci between affected and normal individuals, actual causal polymorphisms, or polymorphisms very tightly linked (and physically near) disease mutations, can be identified. A second strategy examines a random set of SNPs evenly spaced throughout the genome with the hope that population LD will exist between one or more markers and a disease locus that will allow the locus to be identified by comparing the allele frequencies of markers in normal versus affected individuals.[80] This strategy would use the human genome sequence only to physically locate SNPs throughout the genome, whereas the former would use information from an annotated human genome sequence both to obtain a physical map of the SNPs and to maximize the likelihood of finding an association, and minimize the number of markers needed, by taking account of the known locations of genes.

Both the approaches outlined above have recently been criticized as impractical on several grounds.[9] Undiscovered genes, polymorphisms in regulatory regions, or in intron splice sites, and other non-coding polymorphisms that influence gene expression, may be missed by a screen restricted to SNPs in known coding regions. Limited linkage disequilibrium may exist in many regions of the human genome and especially in association with common disease polymorphisms which, due to their increased age by comparison with rare mutations, may have experienced much more recombination with nearby markers.

8.1 Extent of Linkage Disequilibrium in Human Populations

Theoretical predictions concerning the expected extent of LD on disease chromosomes based on a simple neutral genetic model have agreed rather poorly with analyses of LD in actual population samples. For example, using simulation to study recombination in ancestral genealogies underlying a population of chromosomes based on a coalescent process model[83,84] that assumes exchangeable offspring distributions among individuals (neutrality) and either constant size or exponential population growth, Kruglyak[85] predicted that LD surrounding common mutations will typically extend no further than about 3 kilobases (kb) [or about 0.003cM].

An analysis of genome-wide LD on chromosomes of individuals from a sample of unrelated families of European origin [the Centre d'Etude de Polymorphisme Humain (CEPH) repository] revealed LD over spans ranging from 0.10 to 4 cM (roughly 100kb to 4Mb) in many regions.[86] More recently,[87] a study of LD in several genomic regions in a sample of individuals of northern-European descent (from the US population) found that high levels of LD surrounded common alleles, typically extended up to 60kb. The same study found LD over considerably smaller regions for a sample of individuals from the Nigerian population. Presumably this is because, unlike Europeans, the Nigerian population did not experience a recent population bottleneck. It is postulated that a population bottleneck occurred when the ancestors of modern Europeans migrated out of Africa roughly 200 000 years ago.[87] Episodes of population contraction and expansion can generate high levels of populations linkage disequilibrium. Even in the Nigerian population, however, LD extends further (about 5kb, on average) than predicted by Kruglyak[85] based on neutral coalescent theory.

Analyses of LD around the *APOE* polymorphism associated with late-onset Alzheimer disease showed LD extending up to 40 kb.[88,89] Directional selection can greatly increase LD at nearby loci and thus overall levels of LD may be much higher than would be predicted based on a neutral model.[90] This is also true for loci with epistatic interactions.[91] Other important factors ignored in Kruglyak's analysis,[85] such as population subdivision, also tend to increase LD. As a result, Kruglyak's predictions about the extent of LD are probably too conservative to be useful.

Additional empirical studies, and population genetic models with greater realism, are needed to determine how far LD will typically extend on disease chromosomes. This is an important issue because the 3kb window of LD predicted by Kruglyak would require that at least 500 000 SNPs be surveyed in whole genome screens for association based on LD, and this number is probably too large to be a realistic goal with existing technology. With LD extending to 100kb, on average, roughly 15 000 SNPs would be sufficient, and this is feasible using current genotyping methods and pooled samples. Another problem is that actual LD is distributed throughout the genome in a stochastic manner determined by the random forces of genetic drift, migration, and selection. Disease mutations with the same frequency in a population may differ greatly in the extent of LD with markers at similar distances. The result is that some disease susceptibility genes will be readily identifiable using LD at linked markers and others, with markers at a similar map distance, will not.

8.2 Power of Association Studies for Mapping Complex Disease Genes

Even if a high degree of LD exists, or the causal disease loci are included in the SNPs that are directly typed in a study, a final problem remains. If a susceptibility allele has a weak marginal effect, it may be very difficult to detect even with large samples. Many studies have considered the potential power of association methods to detect disease susceptibility genes of weak, or large, effect.[5,42,92] In the remaining section, we summarize the results from one recent theoretical study[42] examining the power of a locus-by-locus association study as a function of the strength of the locus-specific heritability of a disease susceptibility locus, and the extent of LD between a susceptibility locus and a neutral marker.

8.3 Locus-by-Locus Detection of Disease Genes with Small Effects

The dominant strategy being developed for mapping susceptibility loci influencing complex diseases compares marker allele frequencies between unrelated normal and affected individuals locus by locus. This approach allows population samples to be pooled, with the frequencies of particular alleles in each pooled population estimated using quantitative polymerase chain reaction (PCR), or related techniques.[93] An advantage of this approach is that an increase in the number of individuals sampled has little effect on the cost, or effort, of the genotyping. A disadvantage is that pooling prevents the use of multilocus haplotypes, or the use of individual genotypic combinations, reducing the power of LD-based methods and, in particular, their ability to identify genes that have small marginal effects but large effects in a particular genetic background due to epistasis.

For polygenic traits, the phenotype is actually determined by the joint effects of all genes possessed by an individual at all susceptibility loci. By considering each marker in isolation, geneticists are focusing on the marginal distribution. Mathematically speaking, the marginal distribution is obtained by averaging the phenotype over all the possible genotypes, weighted according to their probabilities, at loci other than the locus whose marginal distribution is the focus of interest. If genes interact to create disease phenotypes, the marginal effect of a susceptibility locus may be much smaller than its effect in a particular genetic background. In certain situations, however, with sufficiently large sample sizes, even genes with small marginal effects may potentially be detected in a locus-by-locus study.

Schork et al.,[42] consider a model of a biallelic locus with additive and dominance effects, but no epistasis. The mean values of the phenotypes, given genotypes –, – +, and ++, are $-a$, d, and a, respectively. Additive and total genetic variance (V_G) attributable to the locus is defined as $V_G = V_A + V_D$, where $V_A = 2pq(a - d(p - q))^2$, where $q = 1 - p$. They assume that the phenotypic variance among individuals with a particular genotype is $\sigma^2 = 1$ in all cases. This implicitly assumes that there is no epistasis, otherwise the residual genetic variance would differ among genotypes. They then consider the locus-specific heritabilities

$$H^2 = \frac{V_G}{V_G + 1}$$

and

$$h^2 = \frac{V_A}{V_G + 1}$$

In other words, the contribution to the phenotypic variance of all the remaining loci, and the environment, is assumed to be 1. The broad sense heritability, H^2, is considered to be the proportion of broad-sense heritable variation due to the locus. A range of values of H^2 and h^2 were considered by the authors as well as the effects of different allele frequencies and levels of linkage disequilibrium between a linked marker and the disease mutation on power and expected allele frequency differences between samples of control and affected individuals. We summarize the most important findings here. First, allowing values of H^2 to vary between 0.038 and 0.5, and focusing on the probability that an individual sampled from the upper, or lower, α-percent of the distribution of phenotypes carries the susceptibility allele, the authors found that for this probability to be greater than 80%, a broad sense heritability of at least 0.33 is needed, as well as a frequency of the susceptibility allele of greater than 0.3. Although large, these values are not exceptional and suggest that common disease alleles of moderate effect should be detectable in a locus-by-locus study.

Schork et al.,[42] also examined the expected sample size needed to detect the susceptibility locus with a power of 80%. They considered either a dominant, recessive, or additive model with frequencies of the susceptibility allele ranging from 0.10 to 0.25, and LD between the disease locus and a linked marker locus, as measured by D′ (see definition in [94]), ranging from 0.25 to 0.75. The marker is assumed to have a frequency of 0.25. Fixing the type I error rate to be 0.05, the required sample sizes ranged from 29 to 1297 for a dominant model. Similarly, fixing the type

I error rate to be 0.00001, the required sample sizes ranged from 86 to 3700 for a dominant model, 146 to 24 819 for a recessive model, and 70 to 3415 for an additive model. With the possible exception of the recessive model with type I error of 0.00001, these sample sizes are all quite practical for genotyping studies using pooled samples.

8.4 Postgenome Strategies for LD Mapping and Linkage Analysis

A final use of a human genome sequence is to narrow the candidate region for a disease gene in an analysis using linkage or LD mapping. Given information (from a mutational database) about the frequencies at which disease mutations occur in the introns, exons, nongenic regions, etc., of known disease genes, as well as information about the positions of genes in a region of the genome (from an annotated human genome sequence) one can predict *a priori* the probability that a disease mutation resides in any given location. This information is updated with information from linkage analysis or LD mapping, to predict the posterior probability that a disease mutation lies in any given region (i.e. the probability based on both the marker-based mapping information and the information from mutational databases and an annotated human genome sequence). Rannala and Reeve[95] recently developed a Bayesian method for LD mapping taking account of information from an annotated human genome sequence and a mutational database. For their analysis, they used a mutation database of simple Mendelian disorders, which is probably inappropriate for complex diseases, but as more alleles influencing susceptibility to complex diseases are identified and compiled in mutation databases, it should be possible to develop an appropriate prior probability distribution for genome-based mapping of complex diseases. Preliminary simulation studies suggest that such genome-contextual mapping methods can be highly efficient in reducing the size of a candidate region.[95]

9. Discussion

A susceptibility locus can have a low locus-specific heritability for at least two reasons: it has a small effect relative to other loci (or environmental factors) influencing the disorder; it is in low frequency in a population but has a relatively large effect. Loci in the first class will be of little practical importance when the effect is very small because they are unlikely to lead to a useful therapy, and will have low power for predicting patient risk. Loci in the second class (which are essentially genes influencing simple Mendelian disorders) will be of greater usefulness, both in developing therapies for affected individuals carrying the gene and in predicting patient risk, but because of their low frequency will be of little importance for the population as a whole.

Future studies of the power of methods for mapping complex traits should consider power in relation to importance. For example, the fact that the methods may have low power to detect genes with very small marginal effects should not be of great concern because either these genes will be rare and of relatively large effect (and detectable by standard linkage methods) or they will be common and of such small effect that they have little practical importance for therapy or genetic counseling. One could potentially evaluate the importance of a gene with a given marginal heritability and frequency by considering how the risk of disease for individuals in a population would be modified by substituting a non-disease-associated allele for the susceptibility allele. This would be similar to classical approaches in quantitative genetics, which quantify the influence of a gene by considering the average effect of a gene substitution.[4]

The basic steps involved in a study aimed at mapping genes influencing complex genetic diseases are illustrated in figure 7. The first choice confronting a scientist embarking on a study aimed at mapping genes for a particular complex disease is the study population to use. Perhaps the most significant potential difference in study designs will be whether patients and controls are sampled from a homogeneous, or a heterogeneous, population. As mentioned above, there are potential advantages to studies using homogeneous populations in terms of increased population linkage disequilibrium between markers and disease genes, and reduced genetic heterogeneity of the disease. A potential disadvantage is that often fewer affected individuals can be obtained than would be available from a larger more heterogeneous population. Once populations have been chosen, clinical studies of phenotypes in affected individuals from the populations can potentially reduce genetic heterogeneity by focusing on homogeneous subsets of forms of disease. At this stage, heritability studies are also advisable to evaluate the level of segregating genetic variation for the disease in the chosen populations, and to predict whether one, or more, major disease loci may exist in the populations.

The next step is to decide whether the study will use unrelated cases and controls, or affected and unaffected individuals from families. If unrelated cases and controls are used, genotyping can be carried out either by estimating allele frequencies in cases and controls using pooled samples, or by multi-locus genotyping of each individual case, or control. Sometimes a combination of the two approaches is used with multi-locus genotyping of cases and pooling of individuals to obtain allele frequencies in controls. In the case of a sample from a heterogeneous population, pooling of samples for cases and controls is inadvisable because of the like-

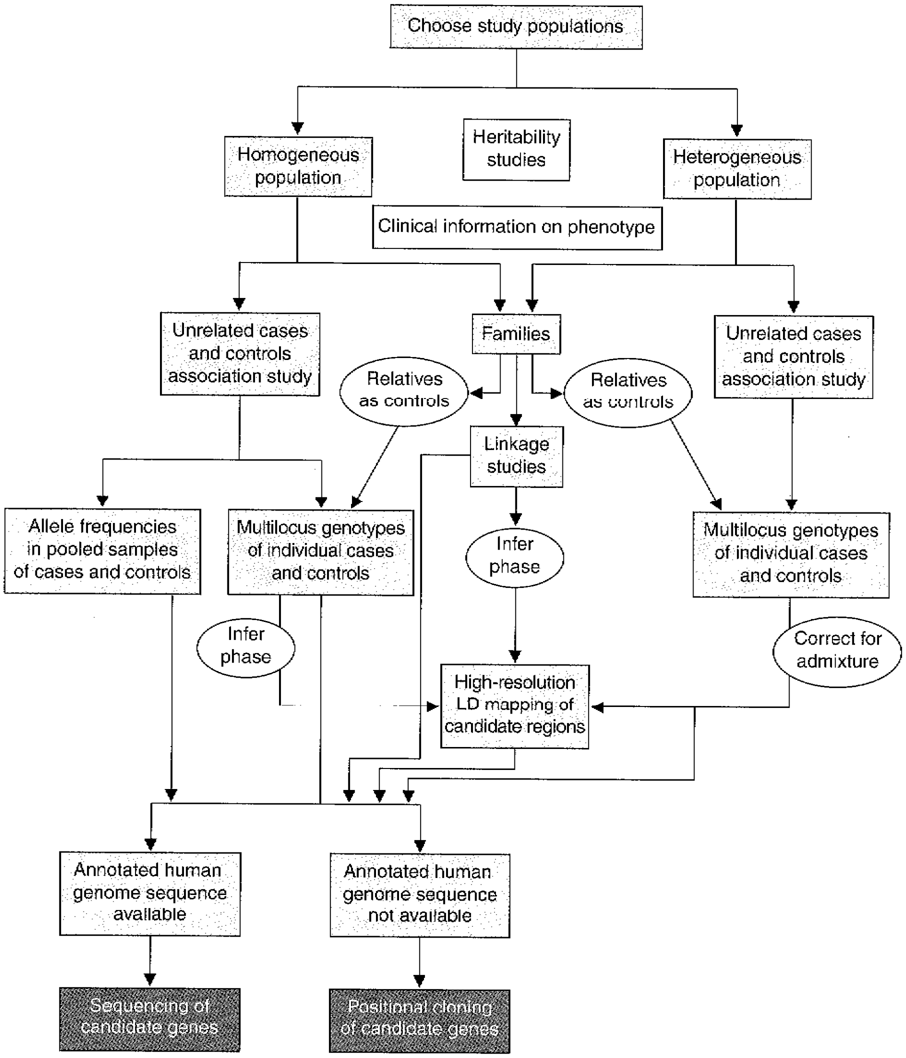
Choose study populations
Homogeneous population
Heritability studies
Heterogeneous population
Clinical information on phenotype
Unrelated cases and controls association study
Families
Unrelated cases and controls association study
Relatives as controls
Relatives as controls
Linkage studies
Allele frequencies in pooled samples of cases and controls
Multilocus genotypes of individual cases and controls
Infer phase
Multilocus genotypes of individual cases and controls
Infer phase
Correct for admixture
High-resolution LD mapping of candidate regions
Annotated human genome sequence available
Annotated human genome sequence not available
Sequencing of candidate genes
Positional cloning of candidate genes

Fig. 7. Flow chart depicting the progression of steps (from top to bottom) involved in an effort to identify genes influencing a complex genetic disease. In many cases, several alternative strategies are possible (see section 9).

lihood that spurious associations may occur due to admixture. In that case, multi-locus genotypes of individuals can be used to take account of population stratification in an association study. If families are used one has a choice of using either linkage methods, association studies using family members as controls for each case, or association studies using unrelated individuals (from different families) as cases and controls. In all three approaches, individual multi-locus genotypes would typically be used.

The final phase in studies using multi-locus genotypes involves either high-resolution mapping to further narrow the candidate regions (genes), or direct analysis of candidate genes in a region identified by either examining an annotated human genome sequence of a candidate region (assuming this exists) or by positional cloning. In practice, LD mapping will often be needed to narrow the number of candidate genes to a sufficiently small number that they can be sequenced in enough affected and control individuals to identify polymorphisms with an increased frequency in affected individuals versus controls.

LD mapping requires that chromosomal haplotypes (phase) be available for linked markers in a candidate region. If families are available, the haplotype phase can often be determined by examining relatives of the individual (from each family) included in the sample of 'unrelated' individuals used for LD mapping.[96] If families are not available, maximum likelihood estimates of haplotype phase can be obtained from population genotype frequencies using an Expectation-Maximization (EM) algorithm,[97] or by calculating the Bayesian posterior probabilities of haplotypes given genotypes,[98] but the methods require some assumptions about population structure (i.e. random mating) that may be unrealistic. If an association study is instead carried out using pooled samples, high-resolution LD mapping will not be possible without further genotyping within the candidate region of a (non-pooled) sample of individuals. Without this second-stage genotyping, one must hope that the associated markers are an actual cause of disease, or are close enough to the causal polymorphisms that these may be identified by sequencing the regions flanking the markers.

In this review, we have provided a brief overview of the stages in a study aimed at mapping genes influencing complex genetic diseases. At each stage, there are many choices as to the way in which the study will be structured. Often, limited information is available about the prospective success of each approach and, at any stage, a negative finding could end the study. Given that both linkage and association methods appear promising for mapping complex disease genes, a conservative strategy would be to sample patients for use in a family-based linkage study and then also use the information available from a sub-sample of the individuals, each from an unrelated family, to carry out an association study. An advantage of this strategy is that haplotype phase can often be determined using family members, potentially increasing the power of association and LD studies.

The next few years will undoubtedly see an explosion of large-sample studies aimed at finding genes associated with complex diseases. This is assured by the rapid expansion of databases of SNP polymorphisms in humans, several of which now contain millions of SNPs mapped throughout the genome (e.g. www.ncbi.nlm.nih.gov/SNP). Commercial efforts (e.g. www.dna.com and www.decode.com) are already underway to collect DNA samples from thousands of individuals affected by complex diseases such as multiple sclerosis and type II diabetes. These samples will be used for large-scale screening of tens of thousands of non-coding, missense, and regulatory polymorphisms throughout the human genome, as well as more modest studies of thousands of candidate loci. New extremely high-throughput genotyping technologies (such as mass-array and microarray genotyping) developed by, and for, industry researchers (e.g. www.sequenom.com and www.affymetrix.com) will aid in this hunt.

Future questions that human geneticists interested in complex disease will need to address include factors such as the cost-benefit ratio of mapping susceptibility genes and how this relates to the mapping effort and available resources. Once we have a few more successes, and it is possible to make informed guesses about the distribution of gene effects among loci (and among alleles within loci) for a typical complex disease, we should be better able to predict the long term benefits of such research; the prospective reduction in disease incidence due to increasingly effective diagnosis and treatment, for example. At that point, it may become clear that knowing the genes for some diseases will not mean that we hold the cure. Other goals achievable through genetics might be equally important, however, such as reducing the population incidence of the disease through preventive screening and drug treatment, etc.

In this paper, we have described some of the statistical tools presently available for mapping genes influencing complex diseases. The field is evolving very rapidly and undoubtedly we will have omitted some techniques that will be of great importance in future studies. Conversely, we will have included others that will turn out to be ineffective. Hopefully, we have provided enough references that motivated readers may discover these shortcomings for themselves. A web page providing links to all the articles cited in this paper, as well as other articles not cited, and statistical resources for studies of complex diseases, can be found at http://rannala.org/complex.html. Readers are invited to submit links and literature references for this page to complex@ rannala.org.

Acknowledgements

Support for this research was provided by NIH grant HG01988 to the author.

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265: 2037-48
2. Vogel F, Motulsky AG. Human genetics: problems and approaches. 3rd ed. New York: Springer, 1997
3. Risch N. Searching for genetic determinance in the New Millenium. Nature 2000; 405: 847-56
4. Falconer DS. Introduction to quantitative genetics. 2nd ed. New York: John Wiley and Sons, Inc., 1981
5. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996; 273: 1516-17
6. Guo S-W. Gene-environment interaction and the mapping of complex traits: Some statistical models and their implications. Hum Hered 2000; 50: 286-303
7. Corder EH, Saunders AM, Strittmatter WJ, et al. Apolipoprotein E4 gene dose and the risk of Alzheimer disease in late-onset families. Science 1993; 261: 921-23
8. Roses AD. A model for susceptibility polymorphisms for complex diseases: Apolipoprotein E and Alzheimer disease. J Neurogenet 1997; 1 (1): 3-11
9. Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? Nat Genet 2000 Oct; 26: 151-7
10. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4th ed. New York: John Wiley and Sons, Inc., 1996
11. Lynch M, Walsh B. Genetics and analysis of quantitative traits. Sunderland, Massachusetts: Sinauer Associates, 1998
12. Darwin C. The origin of species by means of natural selection. London: Murray, 1859
13. Galton F. Natural inheritance. London: Macmillan, 1889

14. Pearson K. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philos Trans R Soc Lond B Biol Sci 1903; A200: 1-66

15. Pearson K. Mathematical contributions to the theory of evolution. XII. On a generalized theory of alternative inheritance, with special reference to Mendel's laws. Philos Trans R Soc Lond B Biol Sci 1904; A203: 53-86

16. Fisher RA. The genetical theory of natural selection. Oxford: Clarendon Press, 1930

17. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans of the R Soc Edinburgh 1918; 52: 399-433

18. Wright S. An analysis of variability of the number of digits in an inbred strain of Guinea pigs. Genetics 1934; 19: 506-36

19. Dempster ER, Lerner IM. Heritability of threshold characters. Genetics 1950; 35: 212

20. Curnow RN, Smith C. Multifactorial models for familial diseases in man. J R Statist Soc 1975 A (Pt 2); 138: 131-169

21. Edwards JH. Familial predisposition in man. Br Med Bull 1969; 25: 58

22. Mendell NR, Elston RC. Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks. Biometrics 1974; 30 (1): 41-57

23. Risch N, Zhang H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 1995; 268: 1584-89

24. Gu C, Todorov AA, Rao DC. Genome screening using extremely discordant and extremely concordant sib pairs. Genet Epidemiol 1997; 14: 791-96

25. Allison DB, Heo M, Schork NJ, et al. Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. Hum Hered 1998; 48: 97-107

26. Slatkin M. Disequilibrium mapping of a quantitative-trait locus in an expanding population. Am J Hum Genet 1999; 64: 1765-73

27. Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 1994; 54: 535-43

28. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998; 62: 1198-211

29. Hopper JL. Variance components for statistical genetics: Applications in medical research to characteristics related to human diseases and health. Stat Methods Med Res 1993; 2: 199-223

30. MacGregor AJ, Snieder H, Schork NJ. Twins. Novel uses to study complex traits and genetic diseases. Trends Genet 2000 Mar; 16 (3): 13-4

31. Lange K. Mathematical and Statistical Methods for Genetic Analysis. Berlin: Springer, 1997

32. Commuzzie AG, Allison DB. The search for human obesity genes. Science 1998; 280: 1374-7

33. Elston RC, Stewart J. A general model for the analysis of pedigree data. Hum Hered 1971; 21: 523-42

34. Morton NE, Maclean CJ. Analysis of family resemblance. 3. Complex segregation of quantitative traits. Am J Hum Genet 1974; 26: 489-503

35. Thompson EA, Lin S, Olshen AB, et al. Monte Carlo analysis on a large pedigree. Genet Epidemiol 1993; 10: 677-82

36. MacLean CJ, Morton NE, Lew R. Analysis of family resolutions. IV. Operational characteristics of segregation analysis. Am J Hum Genet 1975; 27: 365-84

37. Morton NE. Trials of segregation analysis by deterministic and macro simulation. In: Chakravarti A, editor. Human Population Genetics: The Pittsburgh Symposium. New York: Van Nostrand Reinhold, 1984: 83-107

38. Eaves LJ. The resolution of genotypes X environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1984; 1: 215-28

39. Tiret L, Abel L, Rakotovao R. Effect of ignoring genotype-environmental interaction on segregation analysis of quantitative traits. Genet Epidemiol 1993; 10: 581-6

40. Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 1990; 46: 229-41

41. Dupuis J, Brown PO, Siegmund D. Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 1995; 140: 843-85

42. Schork NJ, Nath SK, Fallin D, et al. Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am J Hum Genet 2000; 67: 1208-18

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

44. James JW. Frequency in relatives for an all-or-none trait. Ann Hum Genet 1971; 35: 47-8

45. Stricker C, Fernando RL, Elston RC. Linkage analysis with an alternative formulation for the mixed model of inheritance: The finite polygenic mixed model. Genetics 1995; 141: 1651-6

46. Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 1919; 8: 299-309

47. Bailey NTJ. The mathematical theory of genetic linkage. Oxford: Clarendon Press, 1961

48. Collins A, Frézal J, Teague J, et al. A metric map of humans: 23,500 loci in 850 bands. Proc Natl Acad Sci U S A 1996; 93: 14771-5

49. Ott J. Analysis of human genetic linkage. 3rd ed. Baltimore: The Johns Hopkins University Press, 1999

50. Bell J, Haldane JBS. The linkage between the genes for colour-blindness and haemophilia in man. Proc R Soc Lond B Biol Sci 1937; 123: 119-50

51. Haldane JBS, Smith CAB. A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. Ann Eugen 1947; 14: 10-31

52. Smith CAB. Some comments on the statistical methods used in linkage investigations. Am J Hum Genet 1959; 11: 289-304

53. Boehnke M. Limits of resolution of genetic linkage studies. Am J Hum Genet 1994; 55: 379-90

54. Penrose LS. The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. Ann Eugen 1935; 6: 133-8

55. Knapp M. The affected sib pair method for linkage analysis. In: Pawlowitzki I-H, Edwards JH, Thompson EA, editors. Genetic mapping of disease genes. New York: Academic Press, 1997: 147-57

56. Kruglyak L, Lander E. High-resolution genetic mapping of complex traits. Am J Hum Genet 1995; 56: 1212-23

57. Lander E, Kruglyak L. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat Genet 1995 Nov; 11: 241-7

58. Morton NE. Sequential tests for the detection of linkage. Am J Hum Genet 1955; 7: 277-318

59. Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 1997 Sep; 17: 21-4

60. Thomson G. Significance levels in genome scans. Adv Genet 2001; 42: 475-86

61. Ewens WJ, Spielman RS. The transmission/disequilibrium test: History, subdivision, and admixture. Am J Hum Genet 1995 Aug; 57 (2): 455-64

62. Kaplan NL, Martin ER, Weir BS. Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 1997; 60: 691-702

63. Sheffield VC, Stone EM, Carmy R, et al. Use of isolated inbred human populations for identification of disease genes. Trends Genet 1998; 14: 391-6

64. Wright AF, Carothers AD, Pirastu M. Population choice in mapping genes for complex diseases. Nat Genet 1999; 23 (4): 397-404

65. Boehnke M. A look at linkage disequilibrium. Nat Genet 2000; 25 (3): 246-7

66. de la Chapelle A. Wright FA. Linkage disequilibrium mapping in isolated populations: The example of Finland revisited. Proc Natl Acad Sci U S A 1998; 95: 12416-23

67. Enserink M. Human genetics: Start-up claims piece of Iceland's gene pie. Science 2000; 287: 951

68. Science. A genetic bounty. Science 2000; 288: 1735

69. Rannala B, Mountain J. Detecting immigrants by using multilocus genotypes. Proc Natl Acad Sci U S A 1997; 94: 9197-9201

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

71. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999; 65 (1): 220-8

72. Bacanu SA, Devlin B, Roeder K. The power of genomic control. Am J Hum Genet 2000; 66 (6): 1933-44

73. Lander ES, Botstein D. Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. Cold Spring Harbor Symp Quant Biol 1986b; 51: 49-62

74. Kaplan NL, Hill WG, Weir BF, et al. Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 1995; 56: 18-32

75. Graham J, Thompson EA. Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet 1998; 63: 1517-30

76. Rannala B, Slatkin M. Likelihood of disequilibrium mapping and related problems. Am J Hum Genet 1998; 62: 459-73

77. Hästbacka J, de la Chapelle A, Kaitila I, et al. Linkage disequilibrium mapping in isolated founder populations. Nat Genet 1992; 2: 204-11

78. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921

79. Venter JC, Adams MD, Myers EW, et al. The sequencing of the human genome. Science 2001; 291: 1304-51

80. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: Cataloging human DNA sequence variation. Science 1997; 278: 1580-1

81. Xiong M, Jin L. Combined linkage and linkage disequilibrium mapping for genome screens. Genet Epidemiol 2000; 19 (3): 211-34

82. Wu R, Zeng Z. Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 2001; 157: 899-909

83. Kingman JC. On the genealogy of large populations. Journal of Applied Probability 1982; 19A: 27-41

84. Hudson RR. Gene genealogies and coalescent process. Oxford Surveys in Evolutionary Biology 1990; 7: 1-44

85. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999; 22 (2): 139-44

86. Huttley GA, Smith MW, Carrington M, et al. A scan for linkage disequilibrium across the human genome. Genetics 1999; 152 (4): 1711-22

87. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. Nature 2001; 411: 199-204

88. Martin ER, Gilbert JR, Lai EH, et al. Analysis of association at SNPs in the APOE region. Genomics 2000; 63: 7-12

89. Martin ER, Lai EH, Gilbert JR, et al. SNPing away at complex diseases: Analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 2000; 67 (2): 383-94

90. Kaplan NL, Hudson RR, Langley CH. The 'hitchhiking effect' revisited. Genetics 1989; 123 (4): 887-99

91. Kimura M. A model of a genetic system which leads to closer linkage by natural selection. Evolution; 1956 10: 278-87

92. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A 1999; 96: 15173-7

93. Barcellos LF, Klitz W, Field LL, et al. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 1997; 61: 734-47

94. Lewontin RC. The interaction of selection and linkage. I. Genetic considerations: heterotic models. Genetics 1964; 49: 49-67

95. Rannala B, Reeve JP. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet 2001; 69: 159-78

96. Sobel E, Lange K. Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 1996; 58: 1323-37

97. Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotype data using the Expectation-Maximization algorithm. Heredity 1996; 76: 377-83

98. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68: 978-89

Correspondence and offprints: Dr Bruce Rannala, Department of Medical Genetics, 8-39 Medical Sciences Building, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.
E-mail: brannala@ualberta.ca