Searching for genetic determinants in the new millennium

Neil J. Risch

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA

Human genetics is now at a critical juncture. The molecular methods used successfully to identify the genes underlying rare mendelian syndromes are failing to find the numerous genes causing more common, familial, non-mendelian diseases. With the human genome sequence nearing completion, new opportunities are being presented for unravelling the complex genetic basis of non-mendelian disorders based on large-scale genome-wide studies. Considerable debate has arisen regarding the best approach to take. In this review I discuss these issues, together with suggestions for optimal post-genome strategies.

It is now 135 years since the Bohemian monk Gregor Mendel published the results of his breeding experiments on the garden pea, which initiated the modern era of the study of genetics. In Mendel’s time, the abounding theory of heredity postulated a ‘blending’ of the inherited contributions from the two parents. Mendel’s work clearly showed that such blending did not occur, and led to his conclusion of particulate inheritance (the ‘gene’) and rules of segregation. The relevance of Mendel’s work for human traits was first delineated around the turn of the century by Garrod, who reasoned correctly that similar types of transmission rules explained the ‘inborn errors of metabolism’ typically caused by enzyme deficiencies. At the same time, however, there was another school of thought, primarily emanating from statisticians such as Francis Galton and his student, Karl Pearson. They observed family resemblance for a variety of traits such as anthropometric features and intellectual achievement but they could not discern patterns of inheritance in families that were consistent with mendelian laws. Rather, a ‘blending’-type theory seemed more apt, as children’s phenotypes tended to be, on average, midway between the parents, with some variability. The resolution of this dilemma did not appear until 1918, when Ronald Fisher published his seminal paper describing ‘polygenic’ inheritance. Fisher reconciled the two conflicting schools by recognizing that the critical difference lay in the genetic basis for the variation in the trait being studied.

For the traits Mendel studied, the observed variation was due to a simple difference at a single gene (or locus). On the other hand, for the traits studied by the biometrical school, individual differences were not attributable to different alleles at a single locus. Rather, many different genes, each with allelic variations, contributed to the total observed variability in a trait, with no particular gene having a singly large effect. Thus, an individual phenotype results from the sum total of the effects of all the numerous contributing loci. Furthermore, application of the central limit theorem from statistics implicates a continuous normal distribution in the population for such a trait, similar to what is observed. Thus, the lack of mendelian inheritance patterns for numerous human traits did not require the deconstruction of Mendel’s theory, but rather an extension of it to a more complex scenario that related genes to phenotype. It is clear that Mendel’s success hinged entirely on his selection of single-gene traits, for otherwise the simple rules of inheritance would not have revealed themselves.

The past two decades has witnessed an explosion in both molecular and computational technology, which has enabled the identification of genes for a number of inherited human disorders. These successes have been restricted largely to simple mendelian cases, which are by their nature rare, and although important to the individuals who carry these genes, of limited significance in terms of public health. The promise of the same technology solving the problem of more frequent, non-mendelian familial disorders has largely been unfulfilled. At the same time, at this turn of the millennium, we now find ourselves at the threshold of having the entire human DNA sequence on hand (or at least in silica). It is therefore timely to consider how this new information can best be used in future gene-finding studies, and prospects for success.

The genetic basis of human traits and disease

Critical to the discussion of what approaches are best suited to unravel the genetic basis of traits or disease in the new millennium is a working model of what that basis is likely to entail. So far, we still have a view that primarily reflects the Mendelistic–biometricist dialogue of nearly a century ago. Most human disorders that have been genetically characterized are mendelian, essentially because the extant molecular tools have enabled the identification of these genes by positional cloning (described later), a procedure now described as ‘routine’. By contrast, those disorders or traits for which such approaches have failed are depicted as ‘polygenic’, multifactorial or ‘complex’. Often unwilling to cede to a notion of ‘infinite’ genetic complexity, geneticists refer to these cases as ‘oligogenic’ or ‘multigenic’, implicating a tractable degree of complexity.

If one considers that there are estimated to be approximately 100,000 functional genes in humans and functional variation may exist in any of them, the problem becomes apparent. If the genetic variation that contributes to a trait is due to myriad genes, each of modest effect, the task of identifying those individual contributors becomes monumental. The fact is, however, that gene effects typically come in different sizes, even when there are many of them — at least, this has been the lesson from a lengthy history of model systems. There are several measures of gene effects used by geneticists (Box 1). Many human traits, especially disease
outcomes, show family recurrence patterns that are strongly suggestive of interactions between genes or epistasis, implying the existence of multiple, interacting loci.

Finding genes — a historical perspective
Before the early 1980s, genetic risk factors for a disease or trait could be identified only through direct analysis of candidate genes, usually through association studies. Starting soon after their discovery, blood-group systems such as ABO, MN and Rh were tested directly against an array of human diseases, typically with little replicability. However, after the study of tens of thousands of subjects, it seems that ABO shows consistent, but weak, association with a number of traits involving the gastrointestinal tract1.

Case–control studies
The approach often used for such studies is the case–control design, in which a difference in allele frequency is sought between affected individuals and unrelated unaffected controls. From an epidemiological perspective, a major limitation in this approach is the potential for confounding (that is, spurious association resulting from correlation with the true risk factor) leading to artefactual as opposed to causal associations. In this case, the most likely source of confounding is ethnicity, whereby allele frequencies vary by ethnicity and cases and controls are not adequately matched in terms of ethnicity. Although most investigators would at least attempt coarse matching by major demographic groupings (such as race), substratification within racial groups can still lead to bias. This drawback of traditional case–control designs was recognized early on by Lionel Penrose, who recommended the use of unaffected sibs as controls2. This paradigm, originally applied to ABO and duodenal ulcer3, has seen a resurgence in the past few years4–8. The disadvantage of this design is that sib controls are over-matched to the index cases, leading to a loss of power compared with a well-designed study involving unrelated controls7.

Conventional case–control gene-association studies have a long track record of false-positive results. The high false-positive rate has often been attributed to confounding due to stratification, although this has never been proven. It is more likely that the high false-positive rate results from a low prior probability that the few genes
The notion of numerous factors contributing to a trait, be they genetic or environmental, lends itself naturally to a gaussian (or normal) population distribution for the cumulative effect of those factors. This is the fundamental precept of biometrical genetics. Thus, for continuously measurable traits, it is simplest to characterize gene effects in the context of the gaussian distribution. If we consider a single locus $L$ with two variant alleles $A$ and $a$ with population frequencies $p$ and $q = 1 – p$ respectively, there are two important measures we can define$^{13}$. The first is termed displacement, denoted by $t$, which is the number of standard deviations difference between the mean values of the two homozygotes $AA$ and $aa$ (we assume, for simplicity, that the variance within genotype is the same for each genotype). There is an additional parameter, $d$, representing the mean value of heterozygotes $Aa$ relative to the two homozygotes. Thus, a value of $d = 1$ corresponds to equal means for genotypes $AA$ and $Aa$ (that is, $A$ is dominant), whereas $d = 0$ corresponds to equal means for genotypes $AA$ and $aa$ (that is, $A$ is recessive). A value of $d = 0.5$ corresponds to the heterozygotes being exactly intermediate between the two homozygotes, a situation often described as additive. The second important measure of gene effect is the population variance attributable to segregation of the gene. This is given by $V_L(L) = V_L(A) + V_L(G)$, where $V_L(G) = 2pqd[p(1 – d) + qd]^2$ is the ‘additive’ genetic variance and $V_L(A) = pq^2g^2t^2(0.5)^2$ is the ‘dominance’ genetic variance. The proportion of total variance attributable to locus $A$, which we denote $h^2_A$, is then given by $V_L(A)/V_L(L)$, assuming the variance within genotype to be 1.0. It is important to note that $h^2_A$ is a function of both displacement $t$ and the allele frequency $p$. Thus, a rare gene with large displacement (that is, mendelian) may contribute the same proportion to variance as a common gene with modest displacement (that is, non-mendelian; see Fig. 1). In addition, because $h^2_A$ is a function of $p$, its value can vary from one population to another when $p$ varies, even when the displacement $t$ is the same.

The gaussian model described above has a direct extension to discrete outcomes (for example, affected/unaffected). In this case, the quantitative genetic variable is latent (unobserved) and often described as the genetic ‘liability’. Superimposed on the genetic-liability distribution is a risk function, so that risk of disease also varies, even when the displacement $t$ is the same.

From the perspective of biometrical genetics, epistasis refers to non-additive interactions between gene effects (much as dominance refers to non-additive effects between alleles at a single locus). Thus, the genetic variance underlying a trait can include sources of variance involving interactive effects among any number of loci, and these are termed epistatic variance components. Often, these are segregated into terms based on the number of loci involved in the interaction (for example, two, three or four loci)$^{16,17}$. The identification of restriction-fragment length polymorphism (RFLP) markers$^9$ and subsequently abundant highly polymorphic microsatellite (short tandemly repetitive DNA) loci$^{10,11}$ has led to the mapping of myriad mendelian disease loci. Development of more efficient molecular tools, especially high-throughput DNA sequencing, has enabled the identification of disease loci and their mutations by a process characterized as positional cloning. Naturally occurring mutations are identified on the basis of their chromosomal location by taking advantage of the meiotic process of recombination as manifest in families segregating for the disease. Markers closest to the disease gene show the strongest correlation with disease patterns in families, and typically the tracking of recombination events can narrow the region harbouring a disease gene to between 100 and several thousand kilobases.

The remarkable success of positional cloning rests not simply on the advances observed in molecular technology. It also reflects the enormous power of linkage analysis when applied to mendelian phenotypes — that is, those characterized by a (near) one-to-one correspondence between genotypes at a single locus and the observed phenotype (a glossary of terms is presented in Box 3). In terms of
The challenge of non-mendelian inheritance

As noted above, linkage analysis and positional cloning have had a remarkable track record in leading to the identification of the genes for many mendelian diseases, all within the time span of the past two decades. Several of these genes account for an uncommon subset of generally more common disorders such as breast cancer (BRCA-1 and -2), colon cancer (familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPPC)), Alzheimer’s disease (β-amyloid precursor protein (APP) and presenilin-1 and -2) and diabetes (maturity-onset diabetes of youth (MODY)-1, -2 and -3). These successes have generated a strong sense of optimism in the genetics community that the same approach holds great promise for identifying genes for a range of common, familial disorders, including those without clear mendelian inheritance patterns. But so far the promise has largely been unfulfilled, as numerous such diseases have proven refractive to positional cloning.

The likely explanation for this is related to the century-old debate between Mendelists and biometricians. The gene mutations studied by Mendel, and those more recently discovered by positional cloning, are those with large effect and strong genotype–phenotype correlations. They are effectively the ‘low-hanging fruit’ that are easy to harvest. Now, however, we are left with the great majority of the fruit at the top of the tree with no obvious way to reach it. In genetics terms, these are the numerous genes of smaller effect that are likely to underlie most common, familial traits and diseases in humans—that is, the genes more closely related to the biometrical view of the world. Of course, this sharp distinction is artificial, in that in reality gene effects of all magnitudes exist and depend on the

Table 1 Allele sharing Y for ASPs by displacement t, gene frequency p, prevalence K and heritability H

<table>
<thead>
<tr>
<th>t</th>
<th>p</th>
<th>H1</th>
<th>H = 0.20</th>
<th>H = 0.50</th>
<th>H = 0.80</th>
<th>K = 0.001</th>
<th>H = 0.20</th>
<th>H = 0.50</th>
<th>H = 0.80</th>
<th>K = 0.01</th>
<th>H = 0.20</th>
<th>H = 0.50</th>
<th>H = 0.80</th>
<th>K = 0.10</th>
<th>H = 0.50</th>
<th>H = 0.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.01</td>
<td>0.001</td>
<td>0.501</td>
<td>0.502</td>
<td>0.502</td>
<td>0.502</td>
<td>0.501</td>
<td>0.501</td>
<td>0.501</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>0.01</td>
<td>0.11</td>
<td>0.011</td>
<td>0.522</td>
<td>0.518</td>
<td>0.514</td>
<td>0.512</td>
<td>0.510</td>
<td>0.508</td>
<td>0.505</td>
<td>0.504</td>
<td>0.503</td>
<td>0.501</td>
<td>0.501</td>
<td>0.500</td>
<td>0.502</td>
<td>0.502</td>
</tr>
<tr>
<td>0.30</td>
<td>0.026</td>
<td>0.535</td>
<td>0.529</td>
<td>0.524</td>
<td>0.522</td>
<td>0.521</td>
<td>0.518</td>
<td>0.515</td>
<td>0.510</td>
<td>0.508</td>
<td>0.507</td>
<td>0.506</td>
<td>0.505</td>
<td>0.504</td>
<td>0.504</td>
<td>0.503</td>
</tr>
<tr>
<td>0.70</td>
<td>0.028</td>
<td>0.523</td>
<td>0.519</td>
<td>0.516</td>
<td>0.516</td>
<td>0.515</td>
<td>0.513</td>
<td>0.510</td>
<td>0.508</td>
<td>0.506</td>
<td>0.505</td>
<td>0.504</td>
<td>0.503</td>
<td>0.502</td>
<td>0.502</td>
<td>0.502</td>
</tr>
<tr>
<td>1.00</td>
<td>0.015</td>
<td>0.524</td>
<td>0.516</td>
<td>0.512</td>
<td>0.512</td>
<td>0.510</td>
<td>0.507</td>
<td>0.506</td>
<td>0.503</td>
<td>0.502</td>
<td>0.501</td>
<td>0.501</td>
<td>0.501</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>0.01</td>
<td>0.026</td>
<td>0.535</td>
<td>0.529</td>
<td>0.524</td>
<td>0.522</td>
<td>0.521</td>
<td>0.518</td>
<td>0.515</td>
<td>0.510</td>
<td>0.508</td>
<td>0.507</td>
<td>0.506</td>
<td>0.505</td>
<td>0.504</td>
<td>0.504</td>
<td>0.503</td>
</tr>
<tr>
<td>0.70</td>
<td>0.026</td>
<td>0.523</td>
<td>0.519</td>
<td>0.516</td>
<td>0.516</td>
<td>0.515</td>
<td>0.513</td>
<td>0.510</td>
<td>0.508</td>
<td>0.506</td>
<td>0.505</td>
<td>0.504</td>
<td>0.503</td>
<td>0.502</td>
<td>0.502</td>
<td>0.502</td>
</tr>
<tr>
<td>1.00</td>
<td>0.015</td>
<td>0.524</td>
<td>0.516</td>
<td>0.512</td>
<td>0.512</td>
<td>0.510</td>
<td>0.507</td>
<td>0.506</td>
<td>0.503</td>
<td>0.502</td>
<td>0.501</td>
<td>0.501</td>
<td>0.501</td>
<td>0.500</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>0.01</td>
<td>0.019</td>
<td>0.666</td>
<td>0.604</td>
<td>0.594</td>
<td>0.576</td>
<td>0.554</td>
<td>0.541</td>
<td>0.515</td>
<td>0.512</td>
<td>0.509</td>
<td>0.508</td>
<td>0.507</td>
<td>0.506</td>
<td>0.505</td>
<td>0.505</td>
<td>0.505</td>
</tr>
<tr>
<td>0.10</td>
<td>0.153</td>
<td>0.791</td>
<td>0.765</td>
<td>0.741</td>
<td>0.701</td>
<td>0.676</td>
<td>0.656</td>
<td>0.584</td>
<td>0.572</td>
<td>0.563</td>
<td>0.561</td>
<td>0.559</td>
<td>0.558</td>
<td>0.556</td>
<td>0.556</td>
<td>0.556</td>
</tr>
<tr>
<td>0.30</td>
<td>0.296</td>
<td>0.734</td>
<td>0.720</td>
<td>0.579</td>
<td>0.576</td>
<td>0.573</td>
<td>0.559</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
<td>0.555</td>
</tr>
</tbody>
</table>
trait being studied, but it is also true that the larger the gene effect, the less frequent it is likely to be.

The problem can be given a quantitative interpretation by reverting to the model presented above (Fig. 1, Box 1). For complex diseases, linkage analysis is based on the sharing of alleles identical by descent at a marker locus or loci by affected relatives. For pairs of affected sibs, the most frequently used group, it is straightforward to predict the increase in allele sharing for a fully informative marker at or near the disease locus as a function of the genetic model (Box 2).

The observations in Box 2, Table 1 and Fig. 3 provide perspective on results of linkage screens for numerous disorders over the past decade. So far, all genes first identified by linkage analysis and subsequently positionally cloned are those with low allele frequency and high displacement (that is, mendelian or near mendelian inheritance). These include the genes listed above for breast cancer, colon cancer, familial Alzheimer’s disease and diabetes. By contrast, no genes with moderate or modest displacement, even for rare disorders, have been identified in this way. The literature is now replete with linkage screens for an array of complex disorders such as schizophrenia, manic-depression, autism, asthma, type 1 and type 2 diabetes, multiple sclerosis and lupus, to name but a few. Although many of these studies have reported significant linkage findings, none has led to convincing replication. Typically, independent studies of the same disorder identify maximal evidence at different chromosomal locations. In effect, linkage analysis, traditionally the most reliable of genetic methods when applied to mendelian traits, has proven to be much less reliable a tool for the study of non-mendelian diseases, with a disappointingly high false-positive rate. The likely explanation is that the biological view is closer to reality than the mendelian view for most human traits and diseases.

This does not necessarily mean that no genes underlying non-mendelian traits can be located by linkage analysis. There are several examples of common alleles that have sufficiently large displacement to have been detected by linkage analysis. One example is the role of HLA in type 1 diabetes, where allele sharing by affected sib pairs (ASPs) has been estimated at about 73% (ref. 15). A second example is the role of apolipoprotein E (apoE) in late-onset Alzheimer’s disease, where the ASP allele sharing is estimated at about 60%. Other examples probably exist but have yet to be identified, although the number is likely to be few. Table 1 and Fig. 3 indicate that increasing sample sizes may ultimately improve the odds, but there is clearly a limit. In addition, studying more extreme (and less frequent) phenotypes is helpful provided such cases are also genetically more extreme. However, gene effects with displacements of less than 1 standard deviation (s.d.), which are likely to represent most effects, will rarely be identified this way.

These comments apply equally to quantitative traits studied in humans. Designs that select individuals with extreme phenotypes, both concordant for high or low trait values and extremely discordant for high and low trait values, tend to be the most powerful. But again, only loci with high heritabilities or large displacements can be readily identified by linkage analysis.1,4

Another question relates to whether larger families with many affected individuals would provide better power than smaller families, such as sib pairs. The answer depends on the frequency of the susceptibility allele. For high-frequency alleles, selection of dense

---

### Table 2 Typology of SNPs and their occurrence

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Number (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Coding, non-synonymous, non-conservative</td>
<td>60–100</td>
</tr>
<tr>
<td>II</td>
<td>Coding, non-synonymous, conservative</td>
<td>100–180</td>
</tr>
<tr>
<td>III</td>
<td>Coding, synonymous</td>
<td>200–240</td>
</tr>
<tr>
<td>IV</td>
<td>Non-coding, 5’ UTR</td>
<td>140</td>
</tr>
<tr>
<td>V</td>
<td>Non-coding, 3’ UTR</td>
<td>300</td>
</tr>
<tr>
<td>VI</td>
<td>Other non-coding</td>
<td>&gt; 1,000</td>
</tr>
</tbody>
</table>

---

Computational details for calculating the expected allele sharing for affected sibs have been given previously.1,4 I assume the frequency of the high-risk allele is p, the number of standard deviations between homozygotes (displacement) is t, total heritability is H, heritability due to the tested locus is Hl, and hence residual correlation between sibs is (H – Hl)/2 and the population prevalence of disease is K. Table 1 provides the expected allele sharing Y for a pair of affected sibs as a function of the parameters p, t, H and K. (Hl is derived from p and t as Hl = p(1 – pt)/[1 + pt – pt2].) As can be seen in Table 1, Y increases directly with t, inversely with K, and inversely with H, but has a more complex relationship with p. When the displacement is 0.5, the increase in allele sharing is uniformly minimal. When t = 1.0, the increase in allele sharing is again minimal for common traits (K = 10%), moderate for low-frequency traits (K = 1%) when the allele frequency p lies between 10 and 50% and total heritability is low (< 50%), and sizeable for rare traits (K = 0.1%), especially with moderate allele frequencies (10–50%) and low total heritability (< 50%). When displacement is large (2.0), excess allele sharing is high for infrequent and rare traits with moderate allele frequencies (10–50%), but more modest for high allele frequencies (p = 70%). For common traits (K = 10%), the excess allele sharing is sizeable only for moderate allele frequencies. Examination of the locus-specific heritability Hl, shows a complex pattern. Rare alleles with large displacement create low heritabilities but high allele sharing (for example, p = 0.01, t = 2.0, Hl = 1.9%), whereas common alleles with large displacement generate high heritabilities but lower allele sharing (for example, p = 0.70, t = 2.0, Hl = 29.6%). Thus, genes with low heritability may be identifiable if they are rare and have large displacement but not if they are common with low displacement. Similarly, genes with high heritability should be detectable unless the allele frequency is high.

Further statistical insight can be obtained by examining expected lod scores as a function of allele sharing Y for different sample sizes. Here I assume N completely informative sib pairs, which gives 2N informative identical-by-descent tetrals. If the probability of allele sharing is Y, then the probability of R alleles shared out of 2N is just 

$$R^N(1 − Y)^{2N − R}.$$ 

For R shared alleles the maximum lod score is:

$$\text{Log}_10 R + (2N − R) \log_2 (2N) − 2N \log_2 N$$

The expected maximum lod score (EMLS) is obtained by taking the sum of lod scores weighted by their probability. The EMLS for sharing values of Y ranging from 0.50 to 0.60 for N = 250, 500, 1,000 and 2,000 ASPs is given in Fig. 3. For a genome-wide search, a lod threshold of 3.6 has been recommended.4 As the figure shows, 250 fully informative sib pairs are sufficient to obtain a significant lod score when Y = 0.60; 500 sib pairs can detect Y values > 56.5%; 1,000 sib pairs can detect values > 55%; and 2,000 sib pairs can detect Y values > 53%. Comparison with Table 1 shows that genes with displacement of 0.5 or less cannot be detected even with 2,000 sib pairs. Genes with larger displacement (> 1.0) should be detectable in many circumstances with 1,000 sib pairs provided the disease is not common (K < 1%). Smaller samples (250 sib pairs) have the power to detect genes only with large displacement (t ≥ 2.0) or genes with intermediate displacement (t = 1.0) with intermediate allele frequency (10–50%) for a rare disease (K = 0.1%).
families is likely to increase parental homozygosity at the disease locus and reduce linkage evidence. On the other hand, for rare alleles with large displacement, dense families are usually optimal, because the probability for such a family to be segregating the allele is increased, enhancing the linkage evidence. However, if genome screens of extended pedigrees have been conducted without success, it is reasonable to conclude that rare genes of large effect are unlikely to exist for the trait studied.

Linkage analysis in model systems has actually been far more successful in locating loci with moderate effects for either quantitative traits (quantitative trait loci or QTLs) or disease outcomes than has linkage analysis in humans. There are several reasons for this: (1) inbred strains are often used, which limit the number of loci involved to those that differ between the two strains; (2) rare alleles with large displacement can become fixed in inbred strains subjected to many generations of positive selection; (3) by design, all parents (in an intercross) or half the parents (in a backcross) are heterozygous and thus informative for linkage; and (4) all offspring come from matings of the same phase and thus can be combined into a single large group for analysis. The lack of all of these features in studies of human linkage has probably led to reduced power, but at least some can be addressed by alternate study designs. For example, reducing human genetic variability (item 1 above) is not possible, although focus on certain populations with reduced genetic variation might be beneficial and has been recommended. As described above, rare alleles with large displacement in humans can often be identified by studying dense, extended pedigrees (item 2). Items 3 and 4 above are generally intractable in human linkage studies. The one situation when (3) and (4) apply in humans is when there is linkage disequilibrium (that is, population association) between a marker allele and trait allele. Indeed, when there is complete disequilibrium (or where the trait and marker allele are the same), the human situation becomes directly analogous to the experimental, as individuals from different families can be combined into single groups based on genotype. However, there is still an important difference. In the experimental situation, complete linkage disequilibrium spans the entire length of a chromosome and diminishes only by $\left(1 - \theta^2\right)$ for a marker at recombination fraction $\theta$ away from the trait locus in a single experimental generation. In humans, the amount of disequilibrium between a trait allele and marker allele depends on trait allele homogeneity and is a function of the time since the allele first arose and population demographic history over that time. Typically, disequilibrium spans very short chromosome segments except for rare, recent mutations. Finally, it is important to note that, despite the initial success and power of linkage analysis to locate trait loci in model organisms, even in this case positional cloning of these genes has remained a significant challenge.

**Back to the future-candidate genes**

The disappointing results from linkage studies coupled with a biometrical view of the world has led to the suggestion of alternative approaches to tackling the genetics of non-mendelian diseases, namely reversion to the study of candidate genes on a large scale or high-density genome scans that are dependent on linkage disequilibrium. However, first it is useful to show directly the greater power of detection of gene effects by direct-association (or linkage-disequilibrium) analysis when the involved variant is in hand as opposed to depending on linkage analysis without linkage disequilibrium (Fig. 4). By using an analysis similar to one described previously, ASPs (for linkage) are compared with case–control pairs for association. Parameterizing the effect of the locus in terms of genotype relative risk ($g$) and allele frequency ($p$), for high relative risks ($g \geq 4$) and intermediate allele frequencies ($p = 0.05$–$0.50$) it is realistic to expect linkage analysis to provide statistical evidence for the location of a disease gene. However, for more modest relative risks ($g < 2$), linkage analysis will not provide such evidence except in unrealistically large samples. By contrast, case–control association studies, even using a stringent significance level ($5 \times 10^{-8}$), provide adequate power for genes with relative risks as low as 1.5 (with $p = 0.10$–$0.70$).

**Random SNPs or coding SNPs?**

The suggestion of genome-wide searches for gene effects using large-scale testing of single nucleotide polymorphisms (SNPs), or perhaps more appropriately simple nucleotide polymorphisms (which could include short deletions and insertions and multinucleotide changes as well as single nucleotide substitutions), has led to considerable discussion of the efficiency of different approaches (see review in this issue by Roses, pages 857–865, for a discussion of SNPs). The original suggestion of Risch and Merikangas was to study coding or promoter variants with potential functional significance. Collins et al. subsequently suggested that non-coding or evenly spaced SNPs with high density could be used to track disease loci through linkage disequilibrium. The number of SNPs required for the latter strategy has been the subject of debate, primarily because the extent of linkage disequilibrium in the human genome has not been well studied on a large scale. As opposed to recombination — a biological phenomenon already measured extensively in humans — linkage disequilibrium is a property of populations, and thus depends heavily on their demographic and social histories. Population isolates such as Finns, Ashkenazi Jews and Mennonites have been shown to demonstrate extensive linkage disequilibrium (up to several percent recombination) around rare disease mutations. The degree to which the same will be true for higher-frequency variants is uncertain, although as a general rule the disequilibrium is likely to decline with increasing allele frequency owing to an older coalescence time.

Some researchers have argued that as many as 500,000 evenly spaced SNPs may be required to detect linkage disequilibrium of sufficient magnitude for mapping purposes, even in population isolates, whereas others have argued that founder populations, especially those that have remained small over an extended time period, such as the Saami of Scandinavia or isolated Sardinian populations, would require far fewer SNPs. Although such populations should improve the chances for detecting rare disease alleles (say less than 5% in frequency), owing to greater linkage disequilibrium per base pair, the same is unlikely to be the case for common alleles (greater than 5% in frequency). Furthermore, the power of...


Another complication is that, in contrast to linkage analysis, a density greatly enhances the chance of including a marker in strong association studies but not linkage analysis is DNA pooling, which allele frequencies are examined and compared in a small region does not exclude a significant gene effect in that region. It may be that the SNPs used there are in modest or no disequilibrium with the disease allele, and/or the allele frequencies are divergent. Thus, it seems that in a genome-wide random SNP approach, even at high density, many disease-causing genes would be missed.

Several arguments favour using SNPs in coding and promoter regions rather than random SNPs. First, it is these variants, a priori, that are most likely to be of functional significance and to influence the traits under study. In fact, these are the variants to which random SNP searches are likely to lead. Second, even if not the causative variant in a gene, such SNPs are as likely (or more likely) to be in linkage disequilibrium with the cause as are randomly placed SNPs.

**Typology of SNPs**

If large-scale SNP searches are to become a useful tool for dissecting complex genetic disease, experimental efficiencies need to be brought to bear on the problem. One major efficiency that is possible with association studies but not linkage analysis is DNA pooling, where allele frequencies are examined and compared in a small number of pools rather than a large number of individuals. However, it will still be useful to reduce the number of SNPs studied in a systematic way. Although some have argued for an SNP every kilobase (where n is between 3 and 100), an alternative approach is to prioritize SNPs based on likely functional significance. The past two decades of study of mendelian traits has provided a rational basis on which to classify genomic variation (for example, based on the type and frequency of mutations observed for mendelian traits). Two recent studies that have scanned genes for polymorphism also enable estimation of the number of such SNPs in the human genome. The typology and estimated number of SNPs is provided in Table 2.

Two alleles at different loci that occur together within an individual more often than would be predicted by chance are called population allelic association. A gene with a large displacement, giving rise to a (near) one-to-one correspondence between genotype and phenotype is mendelian. A gene without large displacement, giving rise to a genotype that is not mendelian. Non-mendelian. A gene with a large displacement, giving rise to a (near) one-to-one correspondence between genotype and phenotype.

**Additive variance.** The component of genetic variance due to the additive effects of alleles segregating in the population.

**Coalescence time.** Number of generations to the most recent common ancestor carrying a mutation or DNA variant currently present in a given population.

**Displacement.** The difference in s.d. between mean values for individuals with alternative homozygous genotypes at a given locus.

**Dominance variance.** The component of genetic variance due to non-additive effects of alleles at the same locus.

**Epistasis.** Genetic variance due to non-additive effects of alleles at distinct loci.

**Founder effect.** Coalescence of a mutation or DNA variant in a given population to one of the original population founders or his/her descendant.

**Genetic heterogeneity.** Distinct alleles at the same or different loci that give rise independently to the same genetic disease.

**Genotype relative risk (GRR).** The risk of disease for one genotype at a locus versus another.

**Heritability.** The proportion of population variance in a trait attributable to segregation of a gene or genes. Can be locus-specific or for all loci combined.

**Identity by descent.** Alleles that trace back to a shared ancestor. For sibs, refers to inheritance of the same allele from a given parent.

**Linkage disequilibrium.** Two alleles at different loci that occur together within an individual more often than would be predicted by random chance. Also called population allelic association.

**Mendelian.** A gene with a large displacement, giving rise to a (near) one-to-one correspondence between genotype and phenotype.

**Microsatellite.** A DNA variant due to tandem repetition of a short DNA sequence (usually two to four nucleotides).

**Non-mendelian.** A gene without large displacement, giving rise to significant overlap of genotype distributions and lack of one-to-one correspondence between genotype and phenotype.

**Restriction-fragment length polymorphism (RFLP).** DNA sequence variability leading to cutting or not by a restriction enzyme. Visualized by different patterns of fragment sizes.

**Sibling relative risk.** The disease risk for a sibling of an affected individual compared to the disease risk in the general population.

**Single nucleotide polymorphism (SNP).** DNA sequence variation due to change in a single nucleotide.

**Box 3 Glossary**

**Additive variance**. The component of genetic variance due to the additive effects of alleles segregating in the population.

**Coalescence time**. Number of generations to the most recent common ancestor carrying a mutation or DNA variant currently present in a given population.

**Displacement**. The difference in s.d. between mean values for individuals with alternative homozygous genotypes at a given locus.

**Dominance variance**. The component of genetic variance due to non-additive effects of alleles at the same locus.

**Epistasis**. Genetic variance due to non-additive effects of alleles at distinct loci.

**Founder effect**. Coalescence of a mutation or DNA variant in a given population to one of the original population founders or his/her descendant.

**Genetic heterogeneity**. Distinct alleles at the same or different loci that give rise independently to the same genetic disease.

**Genotype relative risk (GRR)**. The risk of disease for one genotype at a locus versus another.

**Heritability**. The proportion of population variance in a trait attributable to segregation of a gene or genes. Can be locus-specific or for all loci combined.

**Identity by descent**. Alleles that trace back to a shared ancestor. For sibs, refers to inheritance of the same allele from a given parent.

**Linkage disequilibrium**. Two alleles at different loci that occur together within an individual more often than would be predicted by random chance. Also called population allelic association.

**Mendelian**. A gene with a large displacement, giving rise to a (near) one-to-one correspondence between genotype and phenotype.

**Microsatellite**. A DNA variant due to tandem repetition of a short DNA sequence (usually two to four nucleotides).

**Non-mendelian**. A gene without large displacement, giving rise to significant overlap of genotype distributions and lack of one-to-one correspondence between genotype and phenotype.

**Restriction-fragment length polymorphism (RFLP)**. DNA sequence variability leading to cutting or not by a restriction enzyme. Visualized by different patterns of fragment sizes.

**Sibling relative risk**. The disease risk for a sibling of an affected individual compared to the disease risk in the general population.

**Single nucleotide polymorphism (SNP)**. DNA sequence variation due to change in a single nucleotide.
The recent resurgence of association studies using candidate genes has led to much discussion about design issues. The simplest such design is the epidemiological case–control study, contrasting allele frequencies in cases versus controls. As is true for case–control studies generally, confounding is a problem for inferring a causal relationship between a disease and measured risk factor. One approach to deal with confounding is the matched case–control design, where individual controls are matched to cases on potential confounding factors (for example, age and sex) and the matched pairs are then examined individually for the risk factor to see if it occurs more frequently in the case than in its matched control.

From the genetics perspective, the most serious potential confounder is ethnicity. If cases and controls are not ethnically comparable, then differences in allele frequency will emerge at all loci that differentiate these groups whether the alleles are causally related to disease or not (this phenomenon is sometimes known as stratification artefact). One solution to this problem involves the use of relatives as controls and suffers the loss in power (especially with sibs), or population stratification artefact. One solution to this problem is to use a matched case–control design, where controls are ethnically matched to cases. This can in theory be accomplished by focusing on homogenous and randomly mating populations, where cases and controls will presumably be ethnically comparable. However, such populations may be more of a theoretical ideal than a reality, as non-random mating patterns exist in nearly all groups. Nonetheless, association studies in Finland are less likely to be subject to confounding problems than in heterogeneous North American populations.

Another solution to this problem involves the use of relatives as controls for cases. The first such described design proposed the use of unaffected sibs as controls4–8, and this design has recently seen a resurgence of interest35. Designs involving parents as controls have also been proposed4–8. Among these, perhaps the test most similar in spirit to the epidemiological matched case–control analysis is the transmission disequilibrium test35, in which an allele transmitted by a parent to an affected child is matched to the other allele not transmitted from the same parent; MacNemar’s chi-square test of discordance is then applied to the resulting pairs44 (Fig. 5). The two alleles carried by a parent are of necessity ethnically matched, and thus the stratification artefact is eliminated. The same applies to sib controls, whose genotypes are ethnically matched to the cases. But a significant result from a design using parent or sib controls still does not imply a causal relationship between the tested allele and the disease outcome, because linkage disequilibrium with a linked locus (but not an unlinked locus) will also create a positive result. Nevertheless, it does at least indicate a significant gene effect nearby, if not the tested allele itself. The main drawback of using parents or sibs as controls is either unavailability (for example, with parents for a late-onset disease) and loss of power, especially with sibs (as described below).

Whereas the simple case–control design is the mainstay of epidemiology, other family-based approaches are available that are more efficient. In particular, sampling multiplex families, where more than a single individual is affected, can be significantly more efficient than sampling singletons. The increase in efficiency is also a function of the disease allele frequency, and is most pronounced for rarer alleles. Using previously described methods35,17, I have calculated the number of families and total individuals required to detect a gene effect with \( g = 4.0 \) (for the homozygote) and \( g = 2.0 \) (for the heterozygote), assuming a significance level \( \alpha = 5 \times 10^{-8} \) and power \( 1 - \beta = 80\% \). I evaluate two disease allele frequencies, 5% and 20%, and consider designs including one, two or three affected sibs, where the (two) control individuals are either the parents of the sibship, unaffected sibs, or unrelated.

For all designs except sibs, the efficiency is approximately the same when affected and control samples are pooled. For sibs, greater efficiency is possible with individual genotyping44, so those cases (pooled versus not pooled) are evaluated separately. The results are provided in Table 3. Rarer alleles (0.05 versus 0.20) are always more difficult to detect, but the number of subjects required can be reduced substantially by increasing the number affected in the sibship. Using unaffected sibs as controls leads to two to five times the required sample size as using unrelated subjects, depending on the number of affected sibs. Using parents leads to a 40–80% increase, again depending on number of affected sibs. The main conclusion is that if disease-susceptibility alleles are typically low frequency (say \( \leq 20\% \)), multiplex sibships are particularly advantageous; they are also advantageous for more frequent alleles, but the relative advantage is less.

An important remaining question is whether to use parents or sibs as controls and suffer the loss in power (especially with sibs), or use unrelated controls and risk loss of robustness. Population stratification has been invoked numerous times as the cause for an observed high false-positive rate in association studies using candidate genes, yet it has rarely been demonstrated as the culprit38. More likely, it is the lack of a stringent significance level used in such studies that is the problem. If one assumes the prior probability for any particular gene variant to be associated with a disease outcome to be low, most reported significant associations will be false positives.

Table 3 Sample sizes for candidate-gene association studies for different designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Two controls</th>
<th>No. subjects</th>
<th>p = 0.05</th>
<th>No. subjects</th>
<th>p = 0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. affected sibs</td>
<td></td>
<td>No. affected sibs</td>
<td></td>
<td>No. families</td>
<td>No. families</td>
</tr>
<tr>
<td>1</td>
<td>872</td>
<td>2,616</td>
<td>300</td>
<td>900</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1,251</td>
<td>3,753</td>
<td>417</td>
<td>1,251</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1,715</td>
<td>5,145</td>
<td>604</td>
<td>1,812</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2,032</td>
<td>6,096</td>
<td>655</td>
<td>1,965</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>265</td>
<td>1,080</td>
<td>102</td>
<td>406</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>448</td>
<td>1,792</td>
<td>173</td>
<td>692</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>642</td>
<td>2,968</td>
<td>286</td>
<td>1,144</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>992</td>
<td>3,968</td>
<td>361</td>
<td>1,444</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>121</td>
<td>605</td>
<td>52</td>
<td>260</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>218</td>
<td>1,080</td>
<td>101</td>
<td>505</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>314</td>
<td>1,570</td>
<td>177</td>
<td>885</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>605</td>
<td>3,025</td>
<td>258</td>
<td>1,290</td>
<td></td>
</tr>
</tbody>
</table>

NP, not pooled; P, pooled.
An attractive alternative to using family-based controls is to use random or unlinked genetic markers typed in the same cases and controls to determine the extent of possible confounding by ethnicity. In fact, the same markers can also be used to assess the significance of any putative association, or even used to adjust any candidate gene analysis for potential confounding by stratified analysis. Given the proposals for large-scale genotyping, it seems most likely that this approach will ultimately be most efficient.

**Population variation and replication**

As discussed above, rare variants (< 5% frequency) are most likely to be population specific. In some cases, they may be recent in origin and hence specific to a single founder population or less recent and generally found in one major ethnic group (for example, haemochromatosis mutation C282Y found only in Caucasians). These are the variants that are most readily detected by a random SNP linkage-disequilibrium approach, but at the same time potentially least replicable by studying distinct populations. In this case it would be worthwhile to examine the same gene in other populations (or even the same population) for other functional variants that are associated with a similar phenotypic endpoint. Discovery of such alleles provides the strongest evidence for a causal link between the gene and the trait, as is the case with family-specific mutations in mendelian diseases.

Common alleles (> 10% frequency) are more likely to be found globally. If so, a causal association between a candidate SNP and trait outcome should be reproducible in many ethnically diverse populations. However, whereas pan-ethnic replicability provides support for a causal relationship, its absence does not necessarily negate it. It is well known that the same mutation can cause a major disease phenotype in one strain of mouse but no phenotype in a genetically distinct strain. Thus, background factors (genetic and otherwise) differentiating populations can modify the expression of a gene and lead to different levels of association. For example, this seems to be the case for ApoE and Alzheimer’s disease, where the association exists pan-ethnically but is strongest in Caucasians and Asians, and weaker in Hispanics and African Americans.

Another advantage to having an ethnically diverse sample of individuals/families is that patterns of linkage disequilibrium may differ ethnically, helping to resolve causal from non-causal relationships. While populations with high linkage disequilibrium may be useful for initial detection of SNP associations, several different SNPs may be in strong or complete disequilibrium. Populations with lower levels of disequilibrium can help resolve which SNP effect is primary. Generally, Africans appear to have the lowest levels of linkage disequilibrium and hence are likely to be most useful for such analyses. An example is provided by the association of HLA and narcolepsy. In Caucasian and Asian populations, the alleles DR2 and DQB-0602 are equally associated with the disease (and in complete disequilibrium with each other), whereas in Africans there is incomplete disequilibrium between them and DQB-0602 shows the primary effect.

**Conclusions**

As we move into a new millennium, the association of computational and molecular technological developments, including the sequencing of the human genome, is opening up new and unprecedented opportunities for genetics research. It is appropriate to reflect on the accomplishments of the past century and where the new technology is likely to lead us.

As I have indicated, much of the current debate in human genetics regarding approaches to the study of complex diseases can be reflected back onto the century-long debate between the Mendelists view and the biometrician view of the world. Much of the difference in views can be attributed to the traits chosen for study, with Mendelists focusing on those dominated by single-gene effects and the biometricists focusing on continuous, ‘polygenic’ variation. For most common diseases facing humanity, it is likely that the biometrical view is more apt.

The past two decades have witnessed numerous spectacular applications of positional cloning to identify mendelian human disease genes. But the fact is that the same approach is proving limited in identifying the multitude of genes underlying the more common, complex disorders. Even high-density genome scans with evenly spaced SNPs, depending on linkage disequilibrium, are simply an extension of the same reverse-genetics approach.

At this turn of the millennium, with the completion of the human genome project now in sight, we need to consider the full impact of having the entire human DNA sequence. Although the traditional reverse-genetics approaches (linkage and linkage-disequilibrium analysis) may identify a few of the genetic susceptibility agents we seek, I believe a far greater yield will occur by rethinking this problem from a forward-genetics perspective. Identifying all (or most) of the genes in the human genome, as well as identifying and cataloguing the functional variation lying within them, which occurs naturally in the human population, provides opportunities for studying the impact of those variants on phenotypic outcomes of interest. Functional genomics technology involving microarrays and proteomics will provide added insights regarding gene function on the cellular level, improving our ability to predict phenotypic effects of genes at the organismic level. Nevertheless, efficient study designs will still be required, and multiplex families, the mainstay of linkage-based studies, will still be optimal. However, instead of family-based controls, unrelated controls will emerge as a more powerful and efficient approach (especially for analyses based on pooled DNA samples), and robustness will be maintained by studying a large number of independent SNPs. Sampling families of varying ethnicity will also be advantageous from the perspective of enhancing evidence of causality as well as identifying genetic and/or environmental modifying factors.
Despite future developments, it will still be important to view the study of human disease from an epidemiological perspective. Both human genetics and epidemiology are observational as opposed to experimental sciences, and we will never be able to exert the degree of scientific control in studies of human disease that experimentalists can with model systems. Furthermore, we must not lose sight of the numerous non-genetic influences that influence disease risk, and how they interact with host (that is, genetic) factors.

43. Migliori, E. et al. DBZ*0602 and DBZQA*0102 (DQA1) are better markers than DR2 for macrolymphocytes in Caucasian and Black Americans. Sleep 17, 560–567 (1994).