Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease

David Botstein[1] & Neil Risch[1,2]

doi:10.1038/ng1090

The past two decades have witnessed an explosion in the identification, largely by positional cloning, of genes associated with mendelian diseases. The roughly 1,200 genes that have been characterized have clarified our understanding of the molecular basis of human genetic disease. The principles derived from these successes should be applied now to strategies aimed at finding the considerably more elusive genes that underlie complex disease phenotypes. The distribution of types of mutation in mendelian disease genes argues for serious consideration of the early application of a genomic-scale sequence-based approach to association studies and against complete reliance on a positional cloning approach based on a map of anonymous single nucleotide polymorphism haplotypes.

A historical perspective: mendelian diseases

Connecting phenotype with genotype is the fundamental aim of genetics. Determining the DNA sequences that cause specific traits in the intact organism remains particularly difficult in human genetics, for which experimental interventions (mutagenesis, selection, crosses and DNA transformation) are unavailable and in which the phenotypes of interest may be very subtle. Indeed, no general method for connecting even simple mendelian diseases with the DNA of the genes that cause them was available until 1980, when genome-wide linkage analysis using anonymous DNA polymorphisms was first proposed[1]. The human genetic linkage map and the methods and algorithms developed to construct it provided, for the first time, a robust tool for connecting phenotype with DNA.

Over the past decade, about 1,200 genes causing human diseases or traits have been identified, largely by a process that is generally referred to as 'positional cloning'. Classic examples of successful positional cloning include hemochromatosis[2], nail patella syndrome[3] and lactose intolerance[4]. Through positional cloning, genes controlling mendelian traits or diseases are identified and isolated using only knowledge that the phenotype is inherited. Significantly, no knowledge of the biology of the disease or trait, beyond a secure assessment of the phenotype, is required. Identification of the gene leads immediately to knowledge of the relevant protein or proteins and, often for the first time, any understanding of the molecular and physiological basis of the disease phenotype. Along the way, materials and methods for diagnosis are generated and, in favorable cases, progress toward therapies can be rapid.

The early successes in positional cloning included identification of the genes underlying chronic granulomatous disease[5], the X-linked muscular dystrophies[6], cystic fibrosis[7,8], Fanconi anemia[9], ataxia telangiectasia[10] and neurofibromatosis I (ref. 11), as well as those underlying hereditary predispositions to cancer, including retinoblastoma[12], breast cancer[13,14] and polyposis colorectal cancer[15]. The gene for Huntington disease was mapped early[16] but took another decade to clone[17]. Each of these examples illustrates interesting features of the positional cloning process. For each of these diseases, identification of the underlying genes, the encoded proteins and the types of mutation in affected individuals has provided extraordinary explanatory power about the nature of the disease phenotype.

Linkage mapping: the starting point

Positional cloning begins with linkage analysis. Families in which the disease phenotype segregates are analyzed using a group of DNA polymorphisms. The earliest and still most fully documented success in which linkage mapping alone led to the gene was cystic fibrosis in 1989 (refs. 7, 8). At that time, the polymorphic markers were restriction fragment length polymorphisms[1]; subsequently, simple sequence repeats[18,19], allowing greater information content per locus and a more automated technology, were used. Today, such endeavors benefit increasingly from single nucleotide polymorphisms (SNPs), hundreds of thousands of which have been made available through the sequencing of the human genome[20,21].

[1]*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.* [2]*Division of Research, Kaiser Permanente, Oakland, California 94612, USA. Correspondence should be addressed to D.B. (e-mail: Botstein@genome.Stanford.edu).*

The first and foremost prerequisite for successful linkage mapping is a set of families in which both the disease phenotype is segregating and the assessment of the phenotypes has been made with minimal ambiguity. The determination of linkage is fundamentally a statistical process, and uncertainties introduced by confusion about the affected status of members of a cohort under study produce noise in the best case and completely obscure any linkage signal in the worst case. Therefore, it has been for mendelian diseases such as Huntington disease and cystic fibrosis—for which the diagnosis is least ambiguous and there is a near one-to-one correspondence between genotype and phenotype—that linkage mapping has been spectacularly successful and often accomplished with close to the theoretical minimum number of individuals[22]. Where misdiagnoses, heterogeneity, complex inheritance or frequent phenocopies are abundant—particularly when they result in the inclusion of individuals who have a different disease or no disease at all in the affected group—linkage analysis can fail even in very large cohorts.

Sometimes relatively simple and statistically valid remedies are available that can separate the linkage signal from at least some of the noise. An example of this kind of success was the initial finding of linkage for the breast cancer (*BRCA1*) gene, which predisposes to breast cancer[23]. As we now know, *BRCA1* accounts for only a modest fraction of all breast cancer. Supported by prior epidemiological evidence, the investigators noticed that the signal for linkage improved markedly if a subset of individuals with an earlier age of disease onset was chosen. By including this feature of the *BRCA1* phenotype in the final analysis of the complete data set and allowing for locus heterogeneity, it was possible to find robust statistical evidence for the location of *BRCA1* on chromosome 17 (ref. 23), and the subsequent cloning of the gene was based on this information[13].

But such successes have not been uniform. There is a price to be paid for querying many alternative hypotheses on the same data set and using improvement of signal as the supervising variable. The difficulties introduced by heterogeneity, complexity of the mode of inheritance and misdiagnosis regularly lead to dubious linkages. A cautious and appropriate approach to linkage when the initial results are not straightforward is to use a first data set to try to make an appropriate hypothesis (for example, definition of subgroups by age of onset or by clinical features) and then to replicate the study on a new cohort using the predetermined additional phenotypic ideas derived from the first analysis. The only way to have complete confidence in a finding, even for simple mendelian traits, is independent replication.

Of the genes that have been positionally cloned, most (~90%) were originally mapped in families on a single, simple hypothesis of dominant, co-dominant, recessive or X-linked mendelian inheritance. The nature of linkage mapping limits resolution to the order of 1–10 cM; it is uncommon to have studies of this kind involve more than a few hundred meioses because such diseases are usually rare. The standard procedure, after detecting initial linkage using a set of markers spaced about 5–10 cM apart, is to re-examine the same samples with markers spaced more closely in the region of interest. Even if one has an unlimited supply of closely linked, simple sequence repeats or SNPs, the limit of resolution remains the number of meioses in which crossovers might have occurred. In favorable cases (such as cystic fibrosis), the pattern of crossovers in the region of the gene among the cohorts studied leaves only a few predicted genes, all within about 1 cM (~1 Mb), as likely candidates. In less favorable cases, there may be as many as a few hundred predicted genes that might be the relevant disease gene.

A search of three databases, Online Mendelian Inheritance in Man (OMIM), Human Gene Mutation Database (HGMD) and LocusLink, produces estimates of 1,222, 1,163 and 1,338, respectively, for the number of identified human genes that cause disease. These numbers represent only about 3% of the estimated number of genes in the human genome. The known disease genes are likely to have been the easiest ones to find, and there is every reason to believe that the total number of mendelian disease genes is actually much larger, with many more rare ones remaining to be found in the human population. Increasingly, investigators in search of disease genes are facing difficulties arising from rarity of the phenotype, which makes the collection of adequate numbers of meioses challenging; from problems of diagnosis; or from heterogeneity resulting from different genes that cause indistinguishable phenotypes (for example, Fanconi anemia or xeroderma pigmentosum), admixture of sporadic, non-inherited cases (cancers) or complex genetic etiology.

A way to expedite the identification of rare recessive phenotypes is homozygosity mapping[24], in which affected individuals who are relatives of known degree are examined for regions inherited in common from the ancestors that they share. This method is particularly suitable in populations where marriages among close relatives are common, and many of the rare recessive diseases whose genes have not been characterized are found only in such populations. Homozygosity mapping has the advantages that relatively few individuals are required and that genetic heterogeneity, although still a problem, is mitigated by the reality that each set of closely related affected individuals is very likely to carry the identical mutation.

A quick survey of the literature suggests that homozygosity mapping in inbred families is becoming the method of choice for rare recessive diseases. For example, the Fanconi anemias are caused by loss of any of five genes; three were mapped at least in part by homozygosity, taking advantage of the low probability of heterogeneity[25–27]. A similar history applies to no less than five of the genes that cause variants of the Charcot-Marie-Tooth phenotype[28–32]. Altogether, since 1995 there have been nearly 200 published studies in which identity by descent in individuals who are consanguineous to a known degree has been used to map genes causing rare and/or heterogeneous recessive disease phenotypes.

Historically, the first positional cloning successes were aided by genetic information in addition to linkage. For example, cloning of the genes responsible for chronic granulomatous disease and X-linked muscular dystrophy took advantage of chromosomal aberrations that seemed to segregate with each disease phenotype[5,6]. For both diseases, potential genes were enriched in mRNA preparations made by subtraction hybridization methods.

More general biological and physiological information has been used to find disease genes. When individual genes in the relevant region of the genome have been characterized, and the activities of their protein products seem relevant to the physiology of the disease, then these genes become candidates. An outstanding example of the success of the 'well-informed candidate gene' approach has been the systematic cloning of genes in which mutations cause severely high or low blood pressure[33]. Much more often, insight about the connection between the gene product and the disease occurs only after the gene has been cloned. For several inherited diseases (too many to list here), virtually everything that is known about the biological mechanisms underlying the disease phenotype derives from knowledge of the protein affected by the mutations that cause it. This was and remains the basic motivation for positional cloning.

Linkage disequilibrium: getting closer to the disease gene

Often, after finding statistically credible evidence for linkage, neither chromosomal aberrations nor convincing candidate genes are available. For these diseases, there have been notable successes in which linkage disequilibrium (LD) has been used to narrow further the region of the genome in which the disease gene must lie. Like homozygosity mapping, LD methods depend on identity by descent, in this case from a common (typically older) founder mutation.

In effect, when the mutation is carried by affected individuals descended from a single founder, a large multigenerational pedigree is created in which all of the initial generations are missing up to the most current one or two. Thus, numerous meiotic events (that together serve to map the mutation as in conventional linkage analysis) are generated but cannot be observed directly. Nonetheless, the regions of shared DNA among extant mutation carriers have been shortened by a series of historic recombination events, which lead to significant narrowing of the region in which the disease mutation must lie. Linkage disequilibrium has been used successfully in several positional cloning studies involving both population isolates with more recent ancestral mutations (such as diastrophic dysplasia in Finns[34] and torsion dystonia in Ashkenazi Jews[35]), as well as older mutations found more broadly (for example, cystic fibrosis[7], hemochromatosis[2] and Huntington disease[36] in Europeans).

Initially, LD mapping methods depend on allelic associations between single markers (typically microsatellites) and disease, which are compared between affected subjects and suitable controls from the same population. Recessive diseases are straightforward for such analyses, because both chromosomes carried by the affected individual bear the mutant allele and so the marker alleles associated with mutant chromosomes are not ambiguous. For dominant diseases, however, only one of the two alleles observed at the marker is carried on the same chromosome with the mutation, and this introduces noise into the analysis. But if multiple affected individuals from the same family are genotyped, phased mutation-bearing chromosomes can be defined together with the corresponding marker alleles, thereby eliminating the uncertainty. Even knowing the side of the family from which the disease is inherited (without genotyping another affected family member) can resolve the disease-associated marker allele, provided at least one of the parents of an affected individual is also genotyped.

The typical LD analysis involves a comparison of marker allele frequencies between affected and control individuals (or between disease chromosomes and control chromosomes), but in some circumstances additional information can be obtained by simply examining the genotypic distribution of affected individuals without controls. Deviations from Hardy-Weinberg equilibrium at markers in LD with the disease mutation are expected when the disease inheritance model deviates from simple recessive inheritance or from a multiplicative model[2]. For example, the cloning of the gene mutated in hemochromatosis involved genetic heterogeneity, whereby a subset of affected individuals did not carry the chromosome 6p–associated susceptibility allele. Hardy-Weinberg deviation, manifested by homozygote excess at marker loci, was observed for markers in LD with the mutated allele, and indeed such excess was maximal at the location of the mutated allele—a location identical to that defined by the maximization of the standardized allele frequency difference. Similarly, Hardy-Weinberg deviation can be used to finely map dominant disease alleles, for which markers in LD will show heterozygote excess[2,37].

Greater power in fine-mapping is obtained by haplotype analysis, in which all markers are considered simultaneously as haplotypes rather than individually. Haplotype analysis allows the inference of likely historical crossover points, which localize the disease mutation. When numerous markers are typed, it is often possible to infer such crossover points with a high degree of accuracy, because the preserved and non-preserved portions of the mutation-bearing chromosome will be evident[2]. In less favorable situations, however, historic crossover points may be less obvious. For these cases, new computer algorithms based on haplotype analysis have been developed to estimate statistically the likely locations of such crossovers and thus the likely location of the disease mutation[38–42].

The success of LD mapping also depends heavily on the degree of genetic heterogeneity (both allelic and non-allelic) underlying a disease sample. Unless one or a few mutations accounts for most instances of disease, the signal will be too inconsistent to find associations. But some degree of heterogeneity is tolerable, and newer methods that allow for such heterogeneity by the clustering of disease chromosomes have been shown to be effective[42].

Population genetics has had a primary role in the LD mapping of mendelian diseases, because the method relies on population demographic histories and not simply on observed meioses in families. The method has been applied much more frequently for recessive diseases, because recessive alleles are much less subject to negative selection and thus have the potential to increase in frequency through founder and/or bottleneck effects. By contrast, dominant diseases are much more likely to be characterized by heterogeneity, because dominant alleles are exposed directly to negative selection and so do not persist in the population over long periods of time. Dominant diseases characterized by late onset and/or low penetrance, which can allow founder effects to occur, are the exception to this rule. Two such examples include primary torsion dystonia[35] and Huntington disease[36].

In addition, LD mapping is often most successful in population isolates, which generally show less allelic heterogeneity than do more cosmopolitan populations. The extent of LD can be often predicted on the basis of the elapsed time since the founding of the population, because mutation-bearing chromosomes are most likely to coalesce to the time of founding. For example, greater LD is expected (and observed) for mutations found in French Canadians (founding roughly 400 years ago) than for those found in Finns (founding roughly 2,000 years ago). As the amount of family material (that is, the number of directly observed meioses) is often limited by the infrequency of mendelian diseases, LD mapping has proved to be an invaluable tool in positional cloning, because of its tremendous power for fine-mapping.

Role of genomic resources in gene finding

A primary justification for the extraordinary effort and funding that went into the Human Genome Project (HGP) was the expected value of the infrastructure that would be created in the process of identifying and then studying human disease genes. This value can be seen in the positional cloning of hundreds of disease genes. In the early days, the cost of cloning a single gene was around tens of millions of dollars; today, using the methods and materials created by the HGP (including, of course, the sequence itself), a reasonable estimate is about US $100,000—a roughly 100-fold decrease in expenditure.

The most important infrastructure improvements derived from the transition from methods based on the exchange of physical probes to a purely information-based technology began with the introduction of the sequence-tagged site paradigm for mapping[43]. Physical maps now consist of coordinates on the sequence itself.

Before the sequence was available, the maps consisted of probe sequences tied, again by PCR, to radiation hybrid maps[44] and/or to individual clones. Whereas each research group once had to generate its own maps of the regions of interest for positional cloning, now researchers can go to a variety of websites (such as the National Center for Biotechnology Information (NCBI) website; http://www.ncbi.nih.gov) and find tens of thousands of well-characterized and widely used PCR probe sequences that can be used to pinpoint the gene of interest on the genetic or physical maps. Even if only a little bit of sequence is available, the position and sequences for any probes needed for further investigation can be simply obtained from the NCBI or the UCSC Human Genome Browser (http://www.genome.ucsc.edu) websites.

Bioinformatic resources have been essential in facilitating gene identification. Genomic information is cumulative, and the availability of the aggregated knowledge over the Internet is the central success of the HGP. In addition to the maps and sequences, there are mapping and sequencing programs and services that have made previously tedious and expensive processes cheaper and easier. For example, at one time physical mapping was very labor-intensive; subsequently, radiation hybrid localization became a matter of a few dozen PCR reactions, information sent to a website, and the location to high resolution by return e-mail; now, it is a simple 'while-you-wait' query on the Internet.

The accumulation of sequences of organisms other than human, especially the so-called 'model' organisms in which experimentation is easy, has greatly aided both our understanding and further studies of disease genes and their biological roles. The database resources make similarity searching possible such that homologous sequences in different organisms can be identified. More importantly, the model organism databases regularly provide annotation, based on experiments, that contributes valuable information to understanding the role of a newly cloned gene in the disease of interest. In this way, human genetics has been unified with the genetics of all other organisms on the planet. As we describe further below, cross-species sequence comparisons are also an important facilitator in identifying the genes underlying non-mendelian diseases.

Lessons from cloned mendelian genes

The ultimate demonstration that a gene is responsible for a disease phenotype is the identification of mutations. Making a convincing case for causation is never trivial: innocent polymorphisms are an ever-present confounding factor and their ability to confuse is enhanced by the presence of LD. Identifying several different significant mutations in the same gene segregating in unrelated but clinically similar families offers the most convincing evidence for a causal relationship between a gene and disease. Functional and tissue expression studies offer additional support, which is sometimes crucial when only a single or very few mutations are identified.

Important lessons can be learned from a detailed consideration of the characteristics of mutations in mendelian diseases accumulated over the past decade. Online resources such OMIM, HGMD and LocusLink greatly facilitate such examinations. For example, as of June 2002, HGMD[45] listed over 27,000 mutations in 1,222 genes associated with human diseases and traits. These mutations have been classified by the HGMD by type of change. The relative frequencies of the different types of change are summarized in Table 1. In-frame amino acid substitutions, including changes to nonsense codons, are the most frequent type of mutation (59% of the total), whereas deletions account for 22% of all changes and insertions/duplications account for 7%. The remaining large category of splice-site mutations accounts for 10% of all changes. Less than 1% of mutations have been found in regulatory regions.

Table 1 • Relative frequency of types of mutations underlying disease phenotypes*

Change	Number	% of total
deletion	6,085	21.8
insertion/duplication	1,911	6.8
complex rearrangement	512	1.8
repeat variations	38	0.1
missense/nonsense	16,441	58.9
splicing	2,727	9.8
regulatory	213	0.8
total	27,027	100.0

*Data are from the Human Gene Mutation Database (June 2002).

These data provide overwhelming support for the notion that mendelian clinical phenotypes are associated primarily with alterations in the normal coding sequence of proteins. Although regulatory changes may be underrepresented owing to difficulties in their identification, it seems unlikely to us that the distribution in Table 1 would be modified substantially by including large numbers of regulatory changes that have yet to be identified.

Amino acid replacements can be analyzed further by two criteria: the biochemical severity of missense changes and the location and/or context of the altered amino acid in the protein sequence. Although no absolute rules apply, a useful and frequently used guide is the Grantham scale[46], which categorizes codon replacements into classes of increasing chemical dissimilarity between the encoded amino acids: namely, conservative, moderately conservative, moderately radical, radical and 'stop' or nonsense.

There is a clear relationship between the severity of amino acid replacement and the likelihood of clinical observation. As compared with a conservative amino acid substitution, a nonsense change is 9.0 times more likely to present clinically[47]. The corresponding ratios for radical, moderately radical and moderately conservative changes are 3.0, 2.3 and 1.8, respectively. A parallel trend exists for the relative abundance of the different types of amino acid substitution found in SNPs from human gene sequencing experiments as compared with their abundance in pseudogenes (Fig. 1), reflecting evolutionary selection against radical changes[48].

It is not simply the severity of amino acid substitution that is relevant to clinical outcome, but the importance of that particular amino acid in the overall function of the protein. A way to gauge the importance of a specific amino acid residue is by its conservation across species in metazoan evolution[49]. Indeed, Miller and Kumar[49] have shown for seven human disease-associated genes that disease-causing mutations are significantly more likely to occur at amino acid residues that are perfectly preserved in evolution, as compared with non-preserved residues. Thus, the degree of evolutionary conservation (and therefore intolerance to change) of an amino acid is also an important predictor of the likelihood of clinical significance of a particular substitution.

Summarizing the data from seven disease loci[49], we can also observe a trend between the risk of disease outcome and the extent of evolutionary conservation, with disease probability decreasing monotonically with the number of amino acid differences among species (Fig. 2). This trend is also true, but less marked, for milder disease phenotypes such as those caused by type II, III or IV mutations in glucose 6-phosphate dehydrogenase (G6PD; Fig. 2)[49].

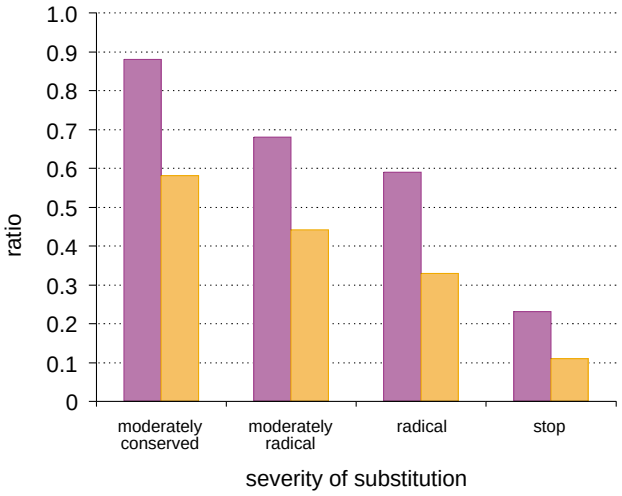

Fig. 1 Clinical severity increases with severity of amino acid substitution. Purple bars represent the ratio of frequencies of the indicated class of change compared to conservative changes for functional human genes compared to pseudogenes; data are derived from ref. 48. Orange bars represent the ratio of the likelihood of clinical observation for a conservative change *versus* the indicated class of change; data are derived from ref. 47.

Another important lesson from the study of mendelian diseases is the strong correlation between clinical severity and severity of the gene lesion. In numerous cases, genotype-phenotype correlation has identified milder forms of disease that are associated with less severe mutations. A classic example is Duchenne (severe) and Becker (mild) muscular dystrophy: originally thought to be genetically distinct, these disease variants were shown to be allelic after the gene was cloned, with Duchenne caused primarily by frame-shift deletions and Becker caused by in-frame changes[50]. Other examples include hemolytic anemias associated with globin mutations[51]; hemochromatosis, with a high penetrance radical amino acid substitution and a lower penetrance milder amino acid substitution[2,52]; Gaucher disease, with a common milder mutation associated with fewer clinical symptoms[53]; and G6PD deficiency, in which the severity of amino acid substitution also correlates with clinical significance[49].

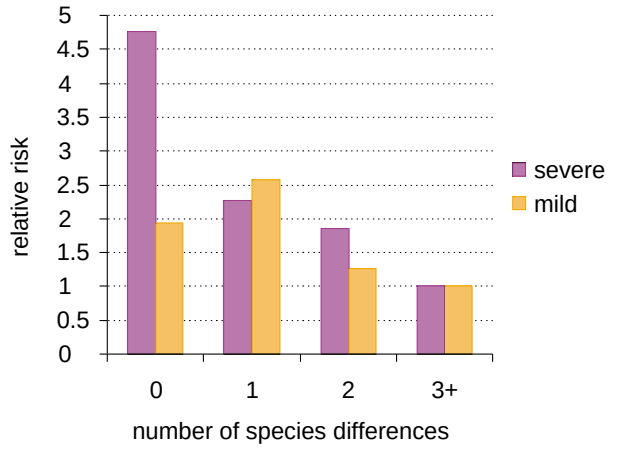

Fig. 2 Evidence that clinical significance correlates with the degree of cross-species evolutionary conservation. Relative risks (approximated by odds ratios) for the observed *versus* the expected number of amino acid changes leading to clinical outcomes are plotted as a function of the number of species differences for that amino acid; three or more differences were used as the reference. Purple bars indicate severe disease mutations (from seven loci); orange bars indicate milder disease mutations (G6PD only). Data are derived from ref. 49.

Notably, for many diseases positional cloning leads first to the identification of the rare, severe, high risk mutations, because these are identified most readily by linkage analysis. Subsequent to cloning, however, higher frequency, milder changes can be also identified, which may be associated (but less strongly) with clinical outcome. In connection with this, a multitude of studies with different experimental organisms support the idea that more radical substitutions produce clearer mutant phenotypes; this notion is supported by an equally large amount of evidence that conservative amino acid substitutions are extremely well tolerated in most proteins. The connection between evolutionary conservation and the likelihood of a phenotype is also supported by extensive studies in experimental model systems (see, for example, ref. 54).

The future: complex diseases

Classical linkage analysis and positional cloning clearly remain the methods of choice for identifying rare, high-risk, disease-associated mutations, owing to the clear inheritance patterns they display. We anticipate that in the future, positional cloning will continue to advance our understanding of the biology of disease phenotypes by identifying the underlying genes. The availability of the human genome sequence, as well as that of model organisms, should expedite this effort. Over the past two decades, many unexpected findings related to mendelian disease loci have appeared that have greatly expanded our view of human biology and physiology. But our current state of knowledge is probably still too limited to predict with any accuracy which gene underlies a particular disease phenotype of interest.

An important lesson has been that 'simple' mendelian inheritance is often not so simple. Multiple different mutations have frequently been identified in the same or in different loci, with variable phenotypic effects and highly variable associated risks. Mutational or genotypic heterogeneity can explain some of the clinical variability observed in single-gene diseases, but usually not all. The residual is probably due to modifier genes and environmental contributors. Identifying such modifiers is a principal challenge for the future, and it is closely allied with resolving in general the genetics of non-mendelian disease, for which all contributing loci might be thought of as 'modifiers' as no single locus of large effect exists. The same arguments apply to diseases in which multi-gene effects and epistasis must be considered.

Although there are likely to be many surprises in the future, we also believe that lessons learned from the identification of over 1,200 human disease genes over the past two decades can be applied to the study of both mendelian and more genetically complex phenotypes. It is clear that, owing to weak linkage signals, positional cloning has limited use in the identification of lower relative risk, disease-associated variants. Attention has therefore turned to association or LD studies in the expectation that these may be more effective[55,56]. Recent large-scale SNP discovery projects have been directed toward this goal; however, no consensus has emerged about the best strategy for identifying complex disease genes[57]. The issues relate to the total number of SNPs in the human genome and the subset of SNPs on which to focus in disease association studies. In the terminology of Peltonen and McKusick[58], the competitive strategies can be categorized as 'map-based' versus 'sequence-based'.

So far, the private sequencing effort has reported 2.1 million unique SNPs[20], and the public SNP consortium has identified 1.4 million[21]. Although each collection contains a modest false-positive rate (10–15%)[59], the false-negative rate is of greater concern. Because neither collection was based on the sequence of a large number of subjects, numerous lower

frequency (<10%) SNPs, especially those that are specific to a single population, were not detected. A more comprehensive sequencing effort (84 ethnically diverse individuals) has been carried out for 313 genes and 720 kb of genomic sequence[48]. Only 2% (or 6% excluding singletons) of the SNPs identified are in dbSNP (the public SNP respository), suggesting that there are many more SNPs than the roughly 1.2 million unique SNPs in this database.

Similarly, a recent sequencing effort of 65% of the unique sequence of chromosome 21 from ten individuals identified 36,000 SNPs (24,000 excluding singletons)[60]. Extrapolation to the whole genome gives a minimum of 6.4 million SNPs. But only 45% of the SNPs in dbSNP were found in this study. Taking into account that only 20 chromosomes were analyzed, one can conclude that the number of SNPs in the human genome (defined by a rare-allele frequency of 1% or greater in at least one population) is likely to be at least 15 million—a number nearly three orders of magnitude greater than the number of genes.

Even with significant breakthroughs in genotyping efficiency, only a modest fraction of all of these SNPs can be genotyped in a reasonably sized association study, and herein lies the debate. Some authors have argued for a map-based approach dependent on LD[60–62]. The underlying philosophy is similar to that of classical positional cloning; that is, no assumption can be made about the types of sequence change underlying disease susceptibility. Because LD seems to have a block-like structure[63,64], it has been suggested that a judiciously chosen subset of SNPs can identify most of the genetic variation in any genomic region through haplotypes. Two recent LD surveys both project, however, that a minimum of 500,000 to 1,000,000 SNPs will need to be genotyped by this approach for subjects of European descent, and perhaps double this number for subjects from African populations[60,62].

The alternative strategy is based on genes and sequence[55,57]. Here, genotyping focuses on SNPs identified in coding regions that alter or terminate amino acid sequence, or (from the data above, less frequently) disrupt splice sites, or (much less frequently) occur in promoter regions. On the basis of several recent gene-based sequencing efforts[48,65,66], we estimate that the total number of such gene-related SNPs in the human genome will be between 50,000 and 100,000 (Table 2). Based on results from cloned mendelian diseases, we can prioritize amino acid replacements according to both the severity of the alteration (Fig. 1) and degree of evolutionary conservation (Fig. 2).

Underlying the discussion of the two approaches to complex disease are several questions, the answers to which will have significant impact on the likelihood of the success of each. These issues involve the allele frequency and the relative risks of alleles associated with complex disease phenotypes, the types of change in sequence that are likely to underlie these alleles, and the best way in which to find them.

Allele frequencies, relative risks

A principal issue revolves around the distribution of frequencies and relative risks for disease-associated alleles[67,68]. The evidence for moderate relative risks underlying complex diseases is probably best demonstrated through the inability of linkage genome scans to produce reliable signals[69]. Although linkage signals are a function of both relative risk and allele frequency[57], modest relative risks are the most likely reason for poor signals, because even very rare mutations are typically identified in mendelian diseases owing to high relative risk.

Random surveys of SNPs have shown consistently, and in accordance with theory, that there is a skewed distribution of allele frequencies, with numbers of SNPs increasing with decreasing allele frequency[48,65,66,70]. Thus, there are many more anonymous SNPs with a minor allele frequency of 1% than with a minor allele frequency of 20%. Functional SNPs—in other words, those occurring in the coding region and leading to (particularly nonconservative) amino acid substitutions or nonsense mutations—are skewed even more towards the lower end of the allele frequency distribution[48,70].

Thus, assumptions about the underlying frequency distribution of disease-associated SNPs depend on the degree of selection that may have acted on them in the past. To a large extent, this will depend on the disease risk associated with the allele and how much the disease impairs reproduction. Thus, early onset, severe diseases may have alleles more skewed toward the lower frequency range than later onset and/or milder diseases. Nonetheless, there are examples of alleles with moderately high population frequency and moderate relative risk for severe, childhood disorders. For example, insulin-dependent diabetes mellitus, which was lethal until modern times, is associated with *HLA DR3* and *DR4* alleles, both of which have substantial relative risks and allele frequencies (in people of European ethnicity). In any event, because of the assumption of lower relative risks, it is not unreasonable to assume an allele frequency distribution that is similar, for example, to SNPs in coding regions identified in large SNP surveys.

Another important issue is the degree to which populations differ in terms of genetic susceptibility and LD structure, and therefore the need to sample different population groups in disease studies. Owing to historical isolation, major racial/ethnic groups differ in allele frequencies[48,71]. These differences are most significant for less common alleles; often a particular allele found in one population at modest frequency is missing from others[48]. Similarly, haplotype structure and LD differ across populations, more so for lower frequency haplotypes. A comprehensive analysis of genetic susceptibility needs to focus on the lower frequency range of potential disease-causing alleles (for example, down to 1%). SNP detection surveys, especially for SNPs in coding sequence, need to include a broad range of racial/ethnic groups.

Population differentiation is also a concern for map-based approaches that focus only on the most common haplotypes

Table 2 • Observed and projected numbers of coding region changes from three studies

	Cargill *et al.*[65]		Halushka *et al.*[66]		Stephens *et al.*[48]	
	114 chromosomes		148 chromosomes		164 chromosomes	
	106 genes		75 genes		313 genes	
Type of change	Observed	Projected	Observed	Projected	Observed	Projected
total amino acid changes	185	52,360	209	83,600	574	55,020
conservative	119	33,680			198	18,980
nonconservative*	66	18,680			376	36,040
I					244	23,390
II					85	8,150
III					38	3,640
IV					9	860
splice-site changes					51	4,890

*Among nonconservative changes, type I is moderately conservative, type II is moderately radical, type III is radical and type IV is nonsense.

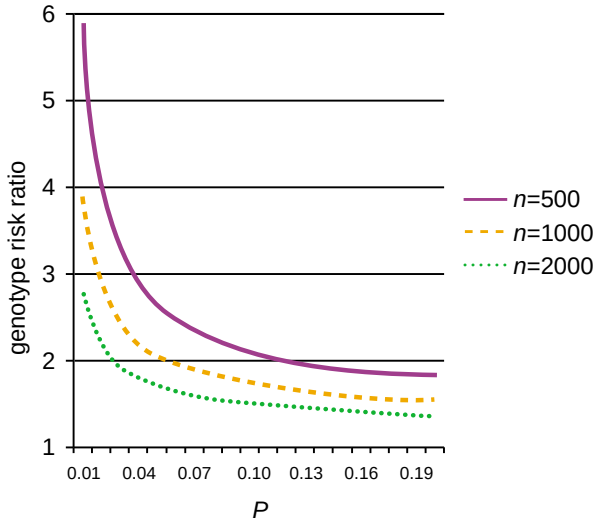

Fig. 3 The genotype risk ratio that is detectable in an association study. Genotype risk ratio is plotted as a function of allele frequency (*P*) for the indicated number (*n*) of case-control pairs.

(those with a frequency of 20% or greater). For example, it is well known that African populations have greater genetic diversity than other populations, and indeed more haplotypes and less LD (that is, shorter 'haplotype blocks')[48,62,72]. For this reason, a haplotype map constructed from European or Asian populations will not be particularly effective in Africans. Similarly, European and Asian populations are dissimilar enough to also lack common sets of SNPs defining haplotype blocks[73].

Another, perhaps more practical approach to the issue of allele frequency is to estimate over what range of values of allele frequency and relative risk could a SNP be identified by an association study, but not by a linkage study. Using formulas described previously[55], we have calculated the values of the genotype risk ratio that could be detected in an association study with varying numbers of case-control pairs (ranging from 500 to 2,000 pairs) as a function of allele frequency (Fig. 3). We have assumed that 50,000 SNPs (100,000 alleles) have been tested for disease association in a genome-wide approach and so use a significance level of 5×10^{-7} and power of 80%. Over this range of values, only unrealistically large linkage studies could produce reliable linkage signals. The genotype risk ratios shown in Fig. 3 seem realistic, even down to allele frequencies of 1%, especially for studies with 1,000 or more pairs. Figure 3 indicates that risk alleles with a frequency of 1% can be readily detected in association studies but not in linkage studies.

A sequence-based approach focusing only on gene regions, which constitute less than 5% of total genomic DNA, should, for the same cost, be able to survey a larger number of individuals to identify SNPs, leading to a much larger number of SNPs in the low frequency range. These SNPs are unlikely to be identified in an association study based on common haplotypes, because the power of such an analysis would be significantly diminished by the discrepancy between the frequency of the disease allele and the associated haplotype[74]. The sample size would need to be increased by a factor of *r*, which is the ratio of the associated haplotype frequency to the disease allele frequency[75]. For example, if the disease allele had a frequency of 3% and the associated haplotype had a frequency of 15%, a sample size five times larger would be required by the haplotype approach.

Can disease-associated alleles be predicted from sequence?

A main feature that distinguishes a map-based approach from a sequence-based approach to genome-wide association studies is the degree to which functional (disease-predisposing) variants can be predicted on the basis of sequence in, for example, coding and/or conserved regions of the genome. The data in Table 1 make a very compelling argument, at least for mendelian phenotypes, in favor of the idea that most diseases are the result of changes that cause loss or alterations in encoded proteins. Less than 1% of the listed mutations occur in regulatory regions, which in general might be more difficult to predict from sequence alone. The greatest risk of a disease phenotype seems to be associated with splice-site mutations, deletions and insertions, because these are the least frequently occurring *a priori*. By contrast, most (60%) of the listed mutations are amino acid substitutions.

A principal concern is the extent to which the distribution observed in Table 1 can be extrapolated to alleles of moderate to low relative risk, which are assumed to underlie complex disease phenotypes. This issue is more difficult to address currently because of the paucity of examples when compared with the mendelian case. The literature does, however, produce a reasonable list of confirmed disease associations for polymorphic alleles with moderate to low relative risk (Table 3). Excluding *HLA*, of the 18 changes listed in Table 3, 1 involves a large deletion in the coding sequence, 1 involves a frameshift mutation, 15 involve amino acid substitutions, and 1 involves variation in the promoter region.

This distribution is not radically different from that shown in Table 1, with the possible exception of a greater proportion of amino acid substitutions. One can draw a similar conclusion from a recent review by Hirschhorn *et al.*[76], who studied a total of 166 putative associations with candidate genes, involving many different sequence changes. Of the six changes that they perceived as soundly confirmed, five were amino acid substitutions and one was a promoter variant. Figure 2 provides further support for the idea that coding region changes are also associated with modest relative risks and milder disease phenotypes. Milder G6PD enzyme deficiencies are associated with amino acid replacements but involve substitutions that show less evolutionary conservation than the severe (hemolytic) mutations[49].

Figure 1 also shows, for mendelian mutations, a correlation between the probability of disease outcome and the severity of amino acid substitution, as measured by the Grantham scale. Clearly, the strongest predictor is nonsense mutation, with a more modest correlation for substitutions. Thus, although the Grantham scale is only partially predictive, in conjunction with evolutionary conservation of the amino acid site its predictive power may be increased. In any event, the number of amino acid substitutions projected for the whole genome (Table 2) is still markedly less than the number of noncoding SNPs and therefore a reasonable collection to test even without prioritization.

Another important issue is the predictive power of expressed sequence tag and cDNA databases, coupled with gene prediction programs, to identify all of the coding regions in the genome. A gene- or sequence-based approach will be only as effective as our ability to identify genes, and current gene-finding strategies are still imperfect. Because of this, the adjunct approach of evolutionary cross-species comparisons[77] may help to identify both gene regions and regulatory regions important for gene function. Focusing on SNP variation in DNA sequence that is highly conserved between species is highly predictive in coding regions (Fig. 2)[49] and may be also predictive outside defined coding regions[77]. Because the number of sequenced species is increasing rapidly,

Table 3 • Some polymorphic, moderate to low risk variants

Disease*	Locus	Change	Frequency	Relative risk
Alzheimer's disease	*APOE–APOE4*	C112R	0.09–0.22	4.0–15.0
	APOE–APOE2	C158R	0.04–0.08	0.5
Thrombosis	factor V Leiden	R506Q	0.00–0.08 (Eur)	5.0–10.0
Hemochromatosis	*Hfe*	H63D	0.02–0.22	4.0
NIDDM	*PPARγ*	P12A	0.85 (Eur)	1.25
IDDM	*INS*	*VNTR* (promoter)	0.85 (Eur)	1.5–2.5
HIV	*CCR5*	Δ32	0.01–0.14 (Eur)	high (resistance), moderate (nonprogression)
Crohn's disease	*NOD2/CARD15*	1007fs	0.02 (Eur)	6.0
		G908R	0.01 (Eur)	6.0
		R702W	0.04 (Eur)	3.0
Breast cancer	*BRCA2*	N372H	0.25 (Eur)	1.3
Colon cancer	*APC*	I1307K	0.03 (AJ)	2.0
Neural tube defects	*MTHFR*	C677T (A→V)	0.30 (Eur)	2.0
		A1298C (E→A)	0.30 (Eur)	2.0
FMF	*MEFV*	P369S	0.02 (AJ)	7.0
		E148Q	0.06 (AJ)	3.0
Graves disease	*CTLA4*	T17A	0.35 (Eur)	1.5–2.0
Creutzfeld-Jakob	*PRNP*	M129V	0.65 (Eur)	3.0
Autoimmune diseases	*HLA B, DR, DQ*	numerous amino acid substitutions	polymorphic	low to moderate

*AJ, Ashkenazi Jews; Eur, European; FMF, familial Mediterranean fever; IDDM, insulin-dependent diabetes mellitus; NIDDM, non-insulin-dependent diabetes mellitus.

the power of this approach may be similarly enhanced. In particular, comparison with the sequence of *Fugu rubripes*[78] might be particularly enlightening because *Fugu rubripes* lacks most of the intronic and repeat sequences found in other higher species.

Numbers of subjects, numbers of SNPs

It seems that a comprehensive map-based approach will require genotyping a much greater number of SNPs than a sequence-based approach—perhaps 10 times as many (50,000–100,000 for sequence-based; 500,000–1,000,000 for map-based). At present this means a marked difference in cost. But advances in technology leading to vastly cheaper genotyping may reduce the impact of this difference, although the cost reduction would have to be substantial because, even in a modest-sized study (500 cases, 500 controls), a minimum of a billion genotypes would be required for the map-based approach versus 50 million for the sequence-based approach.

Should much cheaper genotyping become available, there is an additional statistical issue regarding the much larger number of SNPs typed in a map-based versus a sequence-based approach. For example, the map-based approach requires a significance level of 2.5×10^{-8} to retain the same false-positive rate achieved in the sequence-based approach with a significance level of 5×10^{-7}. To maintain the same power, more subjects would need to be studied by the map-based approach, but the number is only about 20% more[55].

Map-based versus sequence-based: essential differences

Features that distinguish the map-based approach from the sequence-based approach are given in Table 4. In accord with our limited predictive ability relating specific genes to disease, both are agnostic about which particular gene may underlie a disease of interest; however, the sequence-based strategy applies the mendelian experience and the limited available data for modest relative risk alleles to make assumptions about the type and location of putative disease-susceptibility alleles, by focusing on coding and adjacent regions.

To make it practical, haplotyping seems to be crucial for the map-based strategy; this means that every individual must be genotyped separately. By contrast, the sequence-based approach may well permit separate pooling of the DNAs of affected and unaffected individuals as a first-pass approach, providing considerable efficiencies. To capture as many of the SNPs that contribute to phenotypes as possible, we believe that it will be advantageous to collect more, less-frequent SNPs from the coding regions. This will require re-sequencing only about 3% of the total genome of a larger number of individuals to identify the lower frequency SNPs (those in the 1–20% frequency range). Last, and perhaps most important, the number of SNPs to be genotyped in a disease gene search using a sequence-based strategy is likely to be at least tenfold lower than for the map-based

Table 4 • Comparison of genome-wide haplotype map–based *versus* sequence-based strategies

Map-based	Sequence-based
agnostic about gene involved	agnostic about gene involved
agnostic about physical location of functional SNPs	assumes functional SNPs in coding region, splice junctions and promoter regions
agnostic about types of SNPs that are functional	assumes nonconservative changes in conserved amino acids are more likely to be functional
haplotype-based; individual genotyping is usually critical	DNA pooling is possible
detects mostly higher frequency ($P > 0.20$) disease alleles	potential to detect lower frequency disease alleles
detects higher frequency functional SNP soutside coding regions	misses functional noncoding SNPs, except when evolutionarily conserved
requires genotyping 500,000–1,000,000 SNPs or more	requires genotyping 50,000–100,000 SNPs

strategy. The map-based approach provides the opportunity to identify disease-predisposing alleles outside the coding regions of genes, which would be missed by a sequence-based approach, although at much greater expense.

We suggest that an early focus on sequence-based SNPs should be considered, especially in the initial stages of a major program aimed at genome-wide association studies. It makes sense to us to first test the possibility that a select 5% or so of the genome, studied a little more deeply, might yield much of the power needed to find the genes underlying complex diseases.

Conclusion

The past ten years have seen marked achievements in our molecular understanding of human mendelian diseases. We anticipate that the recent deciphering of the human DNA sequence and the variations that lie therein will make the next decade at least as remarkable. We are hopeful that improved genotyping methods and judicious choices regarding design strategies will bring the genetics of complex disease to a point of success comparable to where mendelian genetics now firmly resides.

Acknowledgments

We thank K. Small for technical assistance, and R. Myers and A. Sidow for discussion. This work was supported in part by grants from the National Institutes of Health.

1. Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
2. Feder, J.N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat. Genet.* **13**, 399–408 (1996).
3. Dreyer, S.D. *et al.* Mutations in *LMX1B* cause abnormal skeletal patterning and renal dysplasia in nail patella syndrome. *Nat. Genet.* **19**, 47–50 (1998).
4. Enattah, N.S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
5. Royer-Pokora, B. *et al.* Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature* **322**, 32–38 (1986).
6. Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD). cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509–517 (1987).
7. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
8. Riordan, J.R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
9. Strathdee, C.A., Gavish, H., Shannon, W.R. & Buchwald, M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. *Nature* **356**, 763–767 (1992).
10. Savitsky, K. *et al.* A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* **268**, 1749–1753 (1995).
11. Wallace, M.R. *et al.* Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. *Science* **249**, 181–186 (1990).
12. Fung, Y.-K.T. *et al.* Structural evidence for the authenticity of the human retinoblastoma gene. *Science* **236**, 1657–1661 (1987).
13. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
14. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–792 (1995).
15. Nishisho, I. *et al.* Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665–669 (1991).
16. Gusella, J.F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
17. Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971–983 (1993).
18. Weber, J.L. & May, P.E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
19. Litt, M. & Luty, J.A. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac-muscle actin gene. *Am. J. Hum. Genet.* **44**, 397–401 (1989).
20. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
21. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
22. Lander, E.S. & Botstein, D. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction-fragment-length polymorphisms. *Proc. Natl. Acad. Sci. USA* **83**, 7353–7357 (1986).
23. Hall, J.M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
24. Lander, E.S. & Botstein, D. Homozygosity mapping—a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
25. Gschwend, M. *et al.* A locus for Fanconi anemia on 16q determined by homozygosity mapping. *Am. J. Hum. Genet.* **59**, 377–384 (1996).
26. Saar, K. *et al.* Localisation of a Fanconi anaemia gene to chromosome 9p. *Eur. J. Hum. Genet.* **6**, 501–508 (1998).
27. Waisfisz, Q. *et al.* The Fanconi anemia group E gene, *FANCE*, maps to chromosome 6p. *Am. J. Hum. Genet.* **64**, 1400–1405 (1999).
28. Bolino, A. *et al.* Localization of a gene responsible for autosomal recessive demyelinating neuropathy with focally folded myelin sheaths to chromosome 11q23 by homozygosity mapping and haplotype sharing. *Hum. Mol. Genet.* **5**, 1051–1054 (1996).
29. LeGuern, E. *et al.* Homozygosity mapping of an autosomal recessive form of demyelinating Charcot-Marie-Tooth disease to chromosome 5q23-q33. *Hum. Mol. Genet.* **5**, 1685–1688 (1996).
30. Bouhouche, A. *et al.* A locus for an axonal form of autosomal recessive Charcot-Marie-Tooth disease maps to chromosome 1q21.2q21.3 *Am. J. Hum. Genet.* **65**, 722–727 (1999).
31. Rogers, T. *et al.* A novel locus for autosomal recessive peripheral neuropathy in the EGR2 region on 10q23. *Am. J. Hum. Genet.* **67**, 664–671 (2000).
32. Leal, A. *et al.* A second locus for an axonal form of autosomal recessive Charcot-Marie-Tooth disease maps to chromosome 19q13.3. *Am. J. Hum. Genet.* **68**, 269–274 (2001).
33. Lifton, R.P., Gharavi, A.G. & Geller, D.S. Molecular mechanisms of human hypertension. *Cell* **104**, 545–556 (2001).
34. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations—diastrophic dysplasia in Finland. *Nat. Genet.* **2**, 204–211 (1992).
35. Ozelius, L.J. *et al.* Strong allelic association between the torsion dystonia gene (DYT1) and loci on chromosome 9q34 in Ashkenazi Jews. *Am. J. Hum. Genet.* **50**, 619–628 (1992).
36. MacDonald, M.E. *et al.* The Huntington's disease candidate region exhibits many different haplotypes. *Nat. Genet.* **1**, 99–103 (1992).
37. Klein, C. *et al.* Search for the PARK3 founder haplotype in a large cohort of patients with Parkinson's disease from northern Germany. *Ann. Hum. Genet.* **63**, 285–291 (1999).
38. Service, S.K., Lang, D.W., Freimer, N.B. & Sandkuijl, L.A. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am. J. Hum. Genet.* **64**, 1728–1738 (1999).
39. McPeek, M.S. & Strahs, A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale mapping. *Am. J. Hum. Genet.* **65**, 858–875 (1999).
40. Morris, A.P. & Whittaker, J.C. Fine scale association mapping of disease loci using simplex families. *Ann. Hum. Genet.* **64**, 223–237 (2000).
41. Lam, J.C., Roeder, K. & Devlin, B. Haplotype fine mapping by evolutionary trees. *Am. J. Hum. Genet.* **66**, 659–673 (2000).
42. Liu, J.S. *et al.* Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**, 1716–24 (2001).
43. Brownstein, B.H. *et al.* Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* **244**, 1348–1351 (1989).
44. Cox, D.R. *et al.* Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**, 245–250 (1990).
45. Krawczak, M. *et al.* Human gene mutation database—a biomedical information and research resource. *Hum. Mutat.* **15**, 45–51 (2000).
46. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
47. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring nucleotide effects on the rate of germ-line single base pair substitutions in human genes. *Am. J. Hum. Genet.* **63**, 474–488 (1998).
48. Stephens, J.C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
49. Miller, M.P. & Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10**, 2319–2328 (2001).
50. Gillard, E.F. *et al.* Molecular and phenotypic analysis of patients with deletions within the deletion-rich region of the Duchenne muscular dystrophy (DMD) gene. *Am. J. Hum. Genet.* **45**, 507–520 (1989).
51. Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236 (1979).
52. Risch, N. Haemochromatosis, HFE and genetic complexity. *Nat. Genet.* **17**, 375–376 (1997).
53. Grabowski, G.A. Gaucher disease: gene frequencies and genotype/phenotype correlations. *Genet. Test.* **1**, 5–12 (1997).
54. Palzkill, T. & Botstein, D. Probing β-lactamase structure and function using random replacement mutagenesis. *Proteins Struct. Funct. Genet.* **14**, 29–44 (1992).
55. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
56. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
57. Risch, N. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
58. Peltonen, L. & McKusick, V.A. Dissecting human disease in the postgenomic era. *Science* **291**, 1224–1228 (2001).
59. Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.* **27**, 371–372 (2001).
60. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
61. Collins, F.S., Guyer, M.S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
62. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
63. Daly, M.J. *et al.* High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).
64. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217–222 (2001).
65. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).

66. Halushka, M.K. *et al.* Patterns of single-nucelotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247 (1999).
67. Weiss, K.M. & Terwilliger, J.D. How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26**, 151–157 (2000).
68. Wright, A.F. & Hastie, N.D. Complex genetic diseases: controversy over the Croesus code. *Genome Biol.* **2**, COMMENT 2007 (2001).
69. Altmuller, J. *et al.* Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**, 936–950 (2001).
70. Glatt, C.E. *et al.* Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nat. Genet.* **27**, 435–438 (2001).
71. Dean, M. *et al.* Polymorphic admixture typing in human ethnic populations. *Am. J. Hum. Genet.* **55**, 788–808 (1994).
72. Calafell, F. *et al.* Short tandem repeat polymorphism evolution in humans. *Eur. J. Hum. Genet.* **6**, 38–49 (1998).
73. Osier, M.V. *et al.* A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. *Am. J. Hum. Genet.* **71**, 84–99 (2002).
74. Muller-Myhsok, B. & Abel, L. Genetic analysis of complex diseases. *Science* **275**, 1328–1329 (1997).
75. Risch, N. & Teng, J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
76. Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
77. Sidow, A. Sequence first, ask questions later. *Cell* **111**, 13–16 (2002).
78. Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).