OPINION

Candidate-gene approaches for studying complex genetic traits: practical considerations

Holly K. Tabor, Neil J. Risch and Richard M. Myers

Association studies with candidate genes have been widely used for the study of complex diseases. However, this approach has been criticized because of non-replication of results and limits on its ability to include all possible causative genes and polymorphisms. These challenges have led to pessimism about the candidate-gene approach and about the genetic analysis of complex diseases in general. We believe that these criticisms can be usefully countered with an appeal to the principles of epidemiological investigation.

In the past two decades, many genes that were implicated in simple (Mendelian) diseases have been identified by using genetic linkage and positional cloning methods. Although these methods have been remarkably successful in identifying high RELATIVE RISK genes, they have not been successful in identifying genes that are involved in the complex forms of disease. This failure is the result of three main features of complex diseases. First, such diseases typically vary in severity of symptoms and age of onset, which results in difficulty in defining an appropriate phenotype and selecting the best population to study. Second, they can vary in their aetiological mechanisms, which might involve various biological pathways. Third, and perhaps most importantly, complex diseases are more likely to be caused by several, and even numerous, genes, each with a small overall contribution and relative risk.

Because of these features, researchers have begun to apply other approaches to identify genes that are involved in complex diseases. For example, an association study using a candidate-gene approach looks for a statistical correlation between specific genetic variants and a disease. Association studies are likely to be more effective tools than linkage studies for studying complex traits because they can have greater statistical power to detect several genes of small effect[1]. The candidate-gene approach can be defined as the study of the genetic influences on a complex trait by: generating hypotheses about, and identifying candidate genes that might have a role in, the aetiology of the disease; identifying variants in or near those genes that might either cause a change in the protein or its expression, or be in LINKAGE DISEQUILIBRIUM with functional changes; genotyping the variants in a population; and by using statistical methods to determine whether there is a correlation between those variants and the phenotype.

Rather than rely on markers that are evenly spaced throughout the genome without regard to their function or context in a specific gene, candidate-gene studies focus on genes that are selected because of *a priori* hypotheses about their aetiological role in disease. Furthermore, a candidate-gene study is usually conducted in a population-based sample of affected and unaffected individuals (a case–control study). A candidate-gene study therefore takes advantage of both the increased statistical efficiency of association analysis of complex diseases and the biological understanding of the phenotype, tissues, genes and proteins that are likely to be involved in the disease.

In spite of their promise, candidate-gene studies have been subject to two important criticisms. First, the significant findings of association in many candidate-gene studies have not been replicated when followed up in subsequent association studies. Second, because candidate-gene studies are based on the ability to predict functional candidate genes and variants, some critics argue that current knowledge is insufficient to make these predictions. These critics believe that 'hypothesis-driven' genetic approaches are therefore less likely to yield results than 'anonymous' approaches, which use markers throughout the genome and do not depend on the ability to select biologically plausible candidates.

In this article, we argue that the pessimism that is sometimes aimed at the candidate-gene approach is perhaps too extreme. We suggest that application of the rigorous principles that are used in epidemiology might help respond to some of the criticisms and improve the chances of successfully elucidating genetic components of complex diseases. Although the intersection between candidate-gene and epidemiological approaches has been addressed by other authors, particularly in relation to cancer epidemiology[2], the use of epidemiological principles in the selection, analysis and interpretation of candidate genes and DNA sequence variants has not been described.

Genetic studies and epidemiology

The field of epidemiology is based on observing and measuring disease patterns in populations, and using association and statistical correlation to identify factors that affect those patterns. Epidemiology allows the detection of small-to-moderate, but significant, relative risks of disease that are contributed by several heterogeneous risk

Box 1 | Epidemiological criteria for gene-association studies

- Biological plausibility of association and its consistency with existing knowledge about biology and disease aetiology are evaluated. Is the candidate gene likely to be involved in the phenotype? Are the single-nucleotide polymorphisms (SNPs) likely to have functional effects on the protein?
- The strength of the association between the risk factor and the disease is examined. When considering multiple SNPs in a candidate gene, the ones with strongest association are most likely to be causally related.
- The dose–response relationship of the association is considered. For example, individuals with two copies of a variant might be at greater risk of disease than individuals with one copy of the variant.
- The consistency of the association across past and future studies, and across different populations, is an important consideration. Consistent replication in different populations is strong evidence of causality. Lack of replication does not necessarily imply lack of causality, but might point to the need for more studies in certain populations or more detailed study of the function of a particular gene.

These issues are explored further in REFS 3,8,53.

factors. Similarly, candidate-gene studies aim to detect small-to-moderate relative risks in the context of aetiological and genetic heterogeneity.

Association in epidemiological or candidate-gene studies can be defined as statistical dependence or correlation between two or more events, characteristics or other variables. This dependence is greatly influenced by the characteristics of the study, including the size of the population studied and the number of variables analysed. Findings of association can be influenced by problems such as SELECTION BIAS, RECALL BIAS, MISCLASSIFICATION and CONFOUNDING (for more detailed discussion, see REFS 3–5). Significant associations might be causal, or might simply be the result of coincidence or one of the biases listed above.

Epidemiological principles consider the detection of association in an observational study as a crucial first step in understanding disease aetiology, rather than the key to determining causality. In assessing the causal role of epidemiological associations, epidemiologists use several important guidelines that can also be applied to candidate-gene studies: biological plausibility, strength of association, dose–response relationship and consistency (BOX 1).

From this perspective on association, a candidate-gene study can be seen as a useful first step in exploring potential causal pathways between genetic determinants and complex diseases. Once a statistically significant association is detected, the same gene and genetic variants can be explored in independent populations. In addition, the gene and variant(s) can be studied for association with related phenotypes that might involve similar aetiological pathways. Significant associations can also be used to indicate molecular or biochemical mechanisms for the sequence variants, which allows experiments to be designed to test their functional roles in biological processes and disease pathology. Such studies can provide strong support for causality; for example, in the recent case of identifying a gene that is involved in Crohn disease[6,7].

Criticisms

The criticisms of candidate-gene approaches are rooted in a fundamental challenge to the study of the genetics of complex diseases: in the absence of other effective methods and techniques, how can investigators balance the use of available data about candidate genes and polymorphisms with the desire to minimize the chances of false positives and false negatives? This balance is the focus of epidemiological studies of multiple risk factors and, by reviewing these criticisms from an epidemiological perspective, we hope to provide insight into how this balance can be achieved for candidate-gene studies.

Non-replication. Because findings in association studies are often not replicated in subsequent, independent studies, there is concern that any findings obtained with the candidate-gene approach are unreliable. In a recent article, Ioannidis and colleagues carried out a meta-analysis of 379 studies addressing 36 genetic associations with diseases or traits[8]. This analysis found that association studies of the same disease are often inconsistent in their findings of association, and that the first study to report an association often indicates a stronger effect than is seen in subsequent studies.

There are many reasons for the lack of reproducibility seen with some candidate-gene studies. These reasons indicate caution in both the design and interpretation of such studies, but do not condemn the approach as unreliable. Discrepant findings are often due to variations in study design. For example, candidate-gene studies might differ in the study population and in the definition of the phenotype (examples can be found in REFS 9,10). The same candidate gene or DNA variants might be associated with different relative risks in different populations, and the non-replication might result from real biological differences. Non-replication might also be due to the small magnitude of relative risks that are likely to be detected in candidate-gene studies of complex diseases. The aetiological heterogeneity that is inherent in the label 'complex disease' challenges the measurement and analysis of multiple genetic and environmental components, and the interpretation of multiple and conflicting findings of association[11]. Confounding, bias and misclassification are more likely to obscure small-to-moderate relative risks than larger relative risks.

Another possible explanation for non-replication across candidate-gene studies relates to the selection of polymorphisms that are not likely to be causal. If a polymorphism is selected because of the ease of genotyping and is not likely to affect the function of the protein — such as a restriction fragment length polymorphism in a deep intronic region — then it is assumed, or hoped, that the polymorphism will be in linkage disequilibrium with a 'disease-causing' variant in the gene, and that this will be reflected in a finding of association. Some studies have even genotyped different polymorphisms in the same gene, and assumed that they are in linkage disequilibrium with previously studied polymorphisms without examining the data in their own study population. If this linkage disequilibrium does not exist or varies across populations, different studies might have different findings for the same gene[12]. Therefore, the use of variants that are unlikely to have functional effects and an over-reliance on linkage disequilibrium to detect association might contribute to non-replication of findings of association in candidate-gene studies.

Lack of thoroughness. A second criticism of candidate-gene studies concerns their ability to be thorough and inclusive. Until recently, most researchers either initiated a candidate-gene study after linkage analyses failed to identify genes, or included candidate genes as a peripheral part of a larger epidemiological study that was not originally designed to examine genetic factors[13–15]. Furthermore, most candidate-gene studies considered only a small number of candidate genes and variants. These variants were used simply because they were the only ones that were identified in a gene or because they were easier to genotype than other polymorphisms.

The vast increase in the number of known and putative new genes as a result of the Human Genome Project[16], the identification of many polymorphisms[17] and new high-throughput methods for large-scale genotyping[18] have markedly changed the scope and complexity of the candidate-gene approach. However, the wealth of information has created a need for prioritizing the selection of both the genes and variants. Below, we describe a strategy for prioritizing candidate-gene studies on the basis of probable functional significance. This approach has been used by some groups already, either by selecting only coding single-nucleotide polymorphisms (SNPs) to genotype, or by using more computational approaches to prioritizing SNPs on the basis of their predicted functional effects[19–21]. It balances the desire to maximize the chance of finding a biologically important association with the desire to minimize the chance of detecting false positives or false negatives.

Selecting candidate genes

The selection of candidate genes has many parallels with identifying and ranking risk factors in an epidemiological study. In both arenas, investigators must choose, from a very large number of potential factors, those factors that are most likely to be involved in the phenotype. The first step is usually to examine published studies of the phenotype of interest for suggestions about the types and the number of risk factors, or candidate genes, that are involved. Family and twin studies can be useful in helping to determine the HERITABILITY of a phenotype, the model of inheritance and possibly the penetrance, and a range for the number of genes that are involved. In addition, linkage studies might provide information about genomic regions that can be explored further. These studies can be evaluated from several perspectives, including the population characteristics, the phenotypic definition and the number and density of the markers used[22].

Evidence can also be evaluated for the involvement of specific genes in the phenotype. It is necessary to consider carefully the genes and variants that are selected for these studies and the reasons for their selection. Until recently, a study might have examined a gene solely because of the existence of an easily genotyped polymorphism. Therefore, it is important to determine whether any of the variants that were examined have a functional consequence (see below). In addition, candidate genes that might only have been considered in studies of other phenotypes or that might have not been studied at all can be assessed. Finally, there might be biological, aetiological and pathological models of the disease of interest. Expression studies might provide important information about the tissues and cells that are involved in the disease. For example, by examining expression studies of tumour tissue, investigators have identified genes that might be important in the pathology of certain cancers[23–25]. By examining potential pathogens, carcinogens or environmental factors that are involved in a disease

Box 2 | Database resources for polymorphisms

Even when a variant is identified in the gene of interest using one of the resources described below, it might not be confirmed and there might be little, if any, information about its functional consequences. Some of this information, such as whether the single-nucleotide polymorphism (SNP) results in a missense or nonsense change, can be determined by carefully examining the sequence and structure of the gene in GenBank, if available[54].

In situations in which a candidate gene has been only recently identified, it might be useful to sequence the functional regions of a gene to identify new SNPs. Just as epidemiologists carry out pilot studies to determine population frequencies of certain risk factors of interest, investigators might choose to genotype SNPs of interest in a small representative population to confirm their existence at a useful frequency. Furthermore, because not all genotyping assays work effectively in all genomic contexts, pilot testing can facilitate a smaller-scale evaluation of techniques.

It is also crucial to cross-reference information on SNPs in any of these databases with data in other databases, published literature and the finished sequence of the human genome, as it becomes available. This cross-comparison provides increased confidence in the existence and location of a SNP and its possible functionality.

- **dbSNP**
  The goal of dbSNP is to catalogue variations throughout the genome, regardless of their functional consequences[55]. It is difficult to search dbSNP by gene name. The best way to find SNPs in a gene is to search LocusLink for the gene and select the icon 'V' for variation, which is linked to a dbSNP page that contains all the SNPs in the gene that have been submitted to the database[56].
- **Human Genome Variation Database (HGVbase)**
  HGVbase is focused on documenting genotype–phenotype associations. There is extensive curation and review of polymorphisms before they are entered in the database, and information is included about the genetic location and functional effects of variants. HGVbase can be searched directly with text searches for gene name and can also be searched by sequence[57].
- **The Human Gene Mutation Database (HGMD)**
  HGMD is a database of mutations in the coding regions of human genes that cause inherited disease. The data are obtained through computerized searches of the published literature. HGMD is easier to search than dbSNP and is useful for identifying SNPs that are known to be associated with phenotypes of interest. However, it is less useful for information about newly identified SNPs[58].
- **Disease- and gene-specific databases**
  Over many years, individual laboratories and groups have established and maintained databases that are devoted to cataloguing variations and mutations in specific genes and for specific diseases. These databases vary considerably in their size, scope and degree of quality control. However, they often contain more detailed information about variations within specific well-characterized genes. These databases are listed on the HGMD web site as 'locus-specific mutation databases' and at the HGVbase web site as 'SNP-related databases'.
- **Proprietary databases**
  Among several companies that collect genetic information, Celera has a database that contains information about the SNPs that they identified through the comparison of the Celera human genome sequence to other genome sequence resources. There is substantial overlap between these SNPs and those that are available in dbSNP, as described in the paper of Celera's genome sequence[59]. Genaissance also has a proprietary database, HAP™ Database, that contains information on haplotypes composed of SNPs that were identified by sequencing the functional regions of genes in 93 individuals, and then typing them in individuals from different geographical and disease populations.

process, an investigator might be able to identify genes and proteins that are involved in the processing of these agents. Animal models of a disease process can also provide important information about potential candidate genes and indicate relevant human homologues.

The chosen number of candidate genes and variants, like the number of environmental risk factors, is influenced by many considerations, such as the range of possible hypotheses, the size of the study population and the magnitude of the effect that the investigator hopes to be able to detect with statistical significance[26]. These considerations mirror those involved in the number of risk factors that are selected for evaluation in an epidemiological study.

Prioritizing polymorphisms

A polymorphism is a variation in DNA sequence that has an allele frequency of at least 1% in a population[27]. Approximately 1 in 1,000 bp of the human genome differ between any two chromosome homologues, and studies that compare several individuals in a population or around the world indicate that there are polymorphisms every few hundred base pairs[28,29]. There are several types of polymorphism in the genome: SNPs, repeat polymorphisms and insertions or deletions, ranging from a single base pair to thousands of base pairs in size. Most of the DNA sequence variation in the human genome is in the form of SNPs[17].

Several high-throughput technologies have been used to discover and genotype SNPs[18,30,31] and, with the advent of large-scale sequencing projects, computational methods have been used to identify SNPs in sequence databases[32,33] (see also BOX 2 for a list of database resources for SNPs). In some cases, the SNPs have not been confirmed by sequencing; nevertheless, these databases provide a very large and valuable set of potential SNPs[32].

It is not practical or statistically feasible, at present, to genotype and test all SNPs in the genome for association with phenotypes. Therefore, it is important to select carefully a limited number of SNPs to genotype from the considerable number that are often available in a particular candidate gene. In theory, it is desirable to study only those polymorphisms that affect the function of a protein or its expression, because these are also most likely to affect the risk of a phenotype. However, in most situations, this proof of the effect of polymorphisms on function is not available and is difficult to obtain.

This dilemma is similar to that seen in many epidemiological studies, when a large range of possible risk factors is available for testing and analysis, but the information that is needed to assess the exact biological impact of each risk factor is not known. Epidemiologists use available data to prioritize and select those factors for study that are most likely to be functional and be associated with the risk of disease. Similarly, we believe that it is most effective for investigators who conduct candidate-gene studies to evaluate all possible polymorphisms and prioritize them on the basis of whether they are likely to affect gene function. Those polymorphisms with obvious molecular consequences are more likely to be involved in influencing the risk of disease.

Information about the location and type of the sequence variants in a gene can be used to prioritize polymorphisms (TABLE 1). For some polymorphisms, it might be obvious that a DNA variation changes the function of a protein — for instance, a non-synonymous (missense) variant that alters an amino acid in a protein, or a nonsense change that results in a premature stop codon. These types of polymorphism account for most known disease associations, and therefore they should be given the highest priority for genotyping in candidate-gene studies.

Recently, there has been increased attention on the effects of polymorphisms in transcriptional promoters and regions of DNA that regulate the expression of genes[34]. It is more difficult to predict the effect of a polymorphism in

Table 1 | **Priorities for single-nucleotide-polymorphism selection**

Type of variant	Location	Functional effect	Frequency in genome	Predicted relative risk of phenotype
Nonsense	Coding sequence	Premature termination of amino-acid sequence	Very low	Very high
Missense/ non-synonymous (non-conservative)	Coding sequence	Changes an amino acid in protein to one with different properties	Low	Moderate to very high, depending on location
Missense/ non-synonymous (conservative)	Coding sequence	Changes an amino acid in protein to one with similar properties	Low	Low to very high, depending on location
Insertions/deletions (frameshift)	Coding sequence	Changes the frame of the protein-coding region, usually with very negative consequences for the protein	Low	Very high, depending on location
Insertions/deletions (in frame)	Coding or non-coding	Changes amino-acid sequence	Low	Low to very high
Sense/synonymous	Coding sequence	Does not change the amino acid in the protein — but can alter splicing	Medium	Low to high
Promoter/regulatory region	Promoter, 5′ UTR, 3′ UTR	Does not change the amino acid, but can affect the level, location or timing of gene expression	Low to medium	Low to high
Splice site/intron–exon boundary	Within 10 bp of the exon	Might change the splicing pattern or efficiency of introns	Low	Low to high
Intronic	Deep within introns	No known function, but might affect expression or mRNA stability	Medium	Very low
Intergenic	Non-coding regions between genes	No known function, but might affect expression through enhancer or other mechanisms	High	Very low

UTR, untranslated region.

Table 2 | **Relative risk of functional changes in genes**

Type of change	Relative risk of change
Stop codon	0.13
Radical amino-acid change	0.35
Moderately radical amino-acid change	0.40
Moderately conservative amino-acid change	0.53
Conservative amino-acid change	0.60

The relative risk of each type of change was calculated as an odds ratio using the relative frequency of each type of change in functional genes and pseudogenes, and by comparing them with the relative frequency of silent changes in both types of genes. Relative risks that are closer to zero imply that the type of change is less likely to occur in functional genes than in pseudogenes. Relative risks that are closer to one imply that the kind of change is of roughly equal frequency as silent changes in functional genes and in pseudogenes. Amino-acid changes are categorized on the basis of 'Grantham values', which are derived from physiochemical considerations[60]. Based on data from REF. 42.

a promoter on the basis of the DNA sequence only, but if a DNA sequence variant occurs in a sequence element that is highly conserved in promoters of related genes, it is likely that the polymorphism will have functional consequences. Therefore, it is reasonable to place a high priority on such a sequence variant. It is also possible to test the effects of putative functional polymorphisms in promoters by gene transfection experiments in tissue culture cells; however, this is time consuming and requires appropriate access to laboratory resources and expertise (for example, see REFS 35–37).

Even if a polymorphism in a coding region does not result in an amino-acid change, or if it is not in a coding sequence, it can still affect gene function by altering the stability, splicing or localization of the mRNA[38]. However, except when conserved sequences in splice sites are changed, the effects of non-coding polymorphisms cannot be predicted. In general, synonymous changes are less likely to be associated with disease, and so should be given lower priority for genotyping than coding, promoter/enhancer or splice-site changes. Nevertheless, because of their potential effect on mRNA stability, they should have higher priority than polymorphisms that lie deep within introns[39] (TABLE 1).

There are two key types of data that support this strategy of prioritizing polymorphisms for candidate-gene studies. The first is the evidence from mutational studies of Mendelian diseases. For many diseases, the proportion of cases due to variants in different parts of the gene has been calculated. Although the intronic and regulatory regions of these genes are not always sequenced, most cases are attributable to changes in the coding regions of genes. For example, in Rett syndrome, it is estimated that ~80% of cases are due to changes in the coding regions of the *MECP2* (methyl CpG binding protein 2) gene, most of which are nonsense, missense and frameshift changes[40]. Similarly, coding changes that result in the truncation or absence of protein account for ~80–90% of mutations in *BRCA1* (breast cancer 1, early onset) that are linked to breast cancer[41]. Although variants in genes that are associated with Mendelian traits frequently have severe effects on a protein, it is reasonable to conclude that functional regions of genes are more likely to contain variants that have aetiological effects in complex diseases.

The second type of data that supports this prioritization strategy is from polymorphism discovery studies that have evaluated the kinds of variants and their frequency across hundreds of genes. Stephens *et al.*[42] sequenced the coding regions, regulatory regions and intron–exon boundaries of 313 human genes in 82 individuals of diverse ancestry. The relative frequencies of each type of polymorphism in functional genes compared with pseudogenes are listed in TABLE 2. Radical changes are much less common than less severe changes, presumably because selective pressures reduce the number of changes that affect the function of a protein, but do not do so in pseudogenes. Two other large studies found that variants in coding regions — specifically non-synonymous and nonsense variants, frameshift variants and variants in splice sites — are the least common types of polymorphism[30,31]. However, DNA sequence variants in non-coding regions and coding-region variants that do not change the amino-acid sequence are more frequent in the population. These studies support the contention that it is reasonable to place the highest priority for genotyping on variants that result in changes to the amino-acid sequence, because these variants are most likely to affect the function of the protein, and to be involved in disease aetiology.

In addition to considering the function of polymorphisms, it is also important to consider their frequency in the population to be tested for association. The statistical power to detect a significant association depends on the size of the association and the frequency of the allele of interest[1,26,43]. SNPs with very low allele frequencies would need to have very large relative risks associated with them to be detected in a candidate-gene study, and alleles with very high relative risks would have been detected using linkage analysis. Therefore, SNPs with frequencies of at least 5% are generally more likely to be useful in a candidate-gene study[39]. Because variants with very severe functional consequences are more rare, this might mean placing higher priority on slightly less severe but more common variants. When a SNP is given a high priority on the basis of its position and effect in a gene, it is advantageous to carry out a pilot study to determine its allele frequency in the population that is being tested.

Linkage disequilibrium among SNPs

Another important consideration in selecting SNPs for an association study is whether there is significant linkage disequilibrium in the candidate gene in the study population. Determining linkage disequilibrium in a

Glossary

CONFOUNDING
The distortion of a measure of association, because of the association of other non-intermediate factors with both the variable of interest and the outcome of interest.

HAPLOTYPE
A combination of alleles at different sites on a single chromosome.

HERITABILITY
The proportion of the phenotypic variance due to genetic variance.

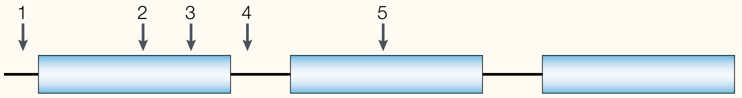
LINKAGE DISEQUILIBRIUM
A population association among alleles at two or more loci. It is a measure of co-segregation of alleles in a population.

MISCLASSIFICATION
Errors in the classification of individuals by phenotype, exposures or genotype that can lead to errors in results. The probability of misclassification can be the same across all groups in a study (non-differential) or vary among groups (differential).

RECALL BIAS
Bias in results due to systematic differences in the accuracy or completeness of recall of past exposures or family history.

RELATIVE RISK
The ratio of the risk of the phenotype among individuals with a particular exposure, genotype or haplotype to the risk among those without that exposure, genotype or haplotype.

SELECTION BIAS
A bias in results due to systematic differences between those who are selected for study and those who are not selected.



SNP1	C/T	Promoter
SNP2	A/G	Non-synonymous
SNP3	A/T	Synonymous
SNP4	T/G	Deep intronic
SNP5	A/C	Non-synonymous

Two major haplotypes SNPs 1–4			
SNP1	SNP2	SNP3	SNP4
C	A	A	T
T	G	T	G

Figure 1 | **Haplotypes and linkage disequilibrium in single-nucleotide-polymorphism selection.** Single-nucleotide polymorphisms (SNPs) 1–4 are in linkage disequilibrium and form two common haplotypes, which can be characterized by any of the variants. Therefore, it is only necessary to determine the genotype of one of the SNPs to capture the information about all four SNPs. SNP5 is not in linkage disequilibrium with SNPs 1–4 and is not part of the haplotype, but it might still contribute independently to the risk of the phenotype.

small pilot sample can help to optimize SNP selection and can provide information about possible HAPLOTYPES for analysis[44]. If several SNPs are in complete linkage disequilibrium in a gene or a region of a gene, it is possible to infer their genotypes on the basis of the genotype of a single SNP. The total number of genotypes assayed can therefore be reduced by selecting the SNPs that are most likely to be functional from a set of SNPs that are in significant linkage disequilibrium (FIG. 1). In some cases, determination of haplotypes or combinations of SNPs that are in linkage disequilibrium might offer more power to detect associations than simply measuring individual SNPs[34,45].

Although an understanding of linkage disequilibrium is a fundamental component of the candidate-gene approach, it is important to note that there is very limited information, at present, about haplotypes within genes in different populations and about linkage-disequilibrium patterns across the genome. However, such data are beginning to be published[42,46]. Therefore, assumptions about the ability to detect associations on the basis of linkage disequilibrium and about the predictive power of haplotypes need to be carefully evaluated[47,48]. For the moment, pilot studies in the relevant population are valuable, but as genome-wide linkage disequilibrium maps, and gene- and population-specific data become available, it might be possible to evaluate potential haplotypes without conducting a pilot study. Discussion of the techniques for predicting haplotypes without DNA information from families and of the relative merits of haplotype analysis in an association study can be found elsewhere[48–50].

In summary, the selection of SNPs for genotyping can be a formidable task in the design of a candidate-gene study. By considering the location and potential function of each SNP, and by evaluating the linkage disequilibrium and potential haplotypes among the SNPs, it is possible to focus on SNPs that are most likely to affect the risk of the phenotype. This requires knowledge about the sequence and structure of the candidate genes and the proteins they encode, and genotyping data on a representative pilot sample of DNA from the study population.

Conclusion

Considerable debate has arisen over the best strategies to use as researchers move forward with SNP-based association studies for the analysis of complex traits[51,52]. Because it is predicted that the human population has many millions of SNPs, it is clear that a prioritization process is necessary. It is unlikely that any one approach will yield all there is to find in the human genome with regard to disease susceptibility. It is our view that, at least for the near future, the use of rigorous epidemiological principles, such as those discussed above, for the choice and analysis of candidate genes and SNPs in disease studies is one tool that might improve the chances of a successful outcome.

Holly K. Tabor is at the Department of Health Research and Policy, and the Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA.

Neil J. Risch is at the Department of Genetics, Stanford University School of Medicine and the Division of Research, Kaiser Permanente, Oakland, California 94611, USA.

Richard M. Myers is at the Department of Genetics, Stanford University School of Medicine.

Correspondence to R.M.M.
e-mail: myers@shgc.stanford.edu

DOI: 10.1038/nrg796

1. Risch, N. J. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Potter, J. D. At the interfaces of epidemiology, genetics and genomics. *Nature Rev. Genet.* **2**, 142–147 (2001).
3. Khoury, M. J., Beaty, T. & Cohen, B. H. *Fundamentals of Genetic Epidemiology* (Oxford Univ. Press, New York, 1993).
4. Hennekens, C. H. & Buring, J. E. *Epidemiology in Medicine* (Little, Brown & Co., Boston, Massachusetts, 1987).
5. Hulley, S. B. *et al. Designing Clinical Research: an Epidemiologic Approach* (Lippincott, Williams & Wilkins, Baltimore, Maryland, 2001).
6. Hugot, J.-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
7. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
8. Ioannidis, J. P. A. *et al.* Replication validity of genetic association studies. *Nature Genet.* **29**, 306–309 (2001).
9. Noble, E. P. The D2 dopamine receptor gene: a review of association studies in alcoholism and phenotypes. *Alcohol* **16**, 33–45 (1998).
10. Palmer, L. J. & Cookson, W. O. Genomic approaches to understanding asthma. *Genome Res.* **10**, 1280–1287 (2000).
11. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
12. Tu, I. P. & Whittemore, A. S. Power of association and linkage tests when the disease alleles are unobserved. *Am. J. Hum. Genet.* **64**, 641–649 (1999).
13. Lindpaintner, K. *et al.* A prospective evaluation of an angiotensin-converting-enzyme gene polymorphism and the risk of ischemic heart disease. *N. Engl. J. Med.* **332**, 706–711 (1995).
14. O'Donnell, C. J. *et al.* Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study. *Circulation* **97**, 1766–1772 (1998).
15. Ensrud, K. E. *et al.* Vitamin D receptor gene polymorphisms and the risk of fractures in older women. For the Study of Osteoporotic Fractures Research Group. *J. Bone Miner. Res.* **14**, 1637–1645 (1999).
16. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
17. The International SNP Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
18. Syvänen, A. Accessing genetic variation: genotyping single-nucleotide polymorphisms. *Nature Rev. Genet.* **2**, 930–942 (2001).
19. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* **12**, 436–446 (2002).
20. Sunyaev, S. *et al.* Prediction of deleterious human alleles. *Hum. Mol. Genet.* **10**, 591–597 (2001).
21. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706 (2001).
22. Terwilliger, J. D. & Goring, H. H. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum. Biol.* **72**, 63–132 (2000).
23. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
24. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
25. Welsh, J. B. *et al.* Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA* **98**, 1176–1181 (2001).
26. Long, A. D. & Langley, C. H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
27. Cavalli-Sforza, L. L. & Bodmer, B. *The Genetics of Human Populations* (W. H. Freeman, San Francisco, California, 1971).
28. Nickerson, D. A. *et al.* Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.* **10**, 1531–1545 (2000).
29. Kwok, P.-Y. *et al.* Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. *Genomics* **31**, 123–136 (1996).
30. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
31. Halushka, M. D. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).

32. Marth, G. T. *et al.* A general approach to single nucleotide polymorphism discovery. *Nature Genet.* **23**, 452–456 (1999).
33. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. & Kwok, P. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8**, 748–754 (1998).
34. Drysdale, C. M. *et al.* Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
35. Myers, R. M., Tilly, K. & Maniatis, T. Fine structure genetic analysis of a β-globin promoter. *Science* **232**, 613–618 (1986).
36. Liu, H. *et al.* Polymorphism in RANTES chemokine promoter affects HIV-1 disease progression. *Proc. Natl Acad. Sci. USA* **96**, 4581–4585 (1999).
37. Theuns, J. *et al.* Genetic variability in the regulatory region of presenillin 1 associated with risk for Alzheimer's disease and variable expression. *Hum. Mol. Genet.* **200**, 325–331 (2000).
38. Cartegni, L., Chew, S. L. & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet.* **3**, 285–298 (2002).
39. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
40. Buyse, I. M. *et al.* Diagnostic testing for Rett syndrome by DHPLC and direct sequencing analysis of the *MECP2* gene: identification of several novel mutations and polymorphisms. *Am. J. Hum. Genet.* **67**, 1428–1436 (2000).
41. Couch, F. J. & Weber, B. L. Mutations and polymorphisms in the familial early-onset breast cancer (*BRCA1*) gene. Breast Cancer Information Core. *Hum. Mutat.* **8**, 8–18 (1996).
42. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
43. Lalouel, J. M. & Rohrwasser, A. Power and replication in case–control studies. *Am. J. Hypertens.* **15**, 201–205 (2002).
44. Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120**, 849–852 (1988).
45. Subrahmanyan, L. *et al.* Sequence variation and linkage disequilibrium in the human T-cell receptor β (*TCRB*) locus. *Am. J. Hum. Genet.* **69**, 381–395 (2001).
46. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
47. Goddard, K. A. B. *et al.* Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
48. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
49. Long, J. C., Williams, R. C. & Urbanek, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810 (1995).
50. Zhao, J. H., Curtis, D. & Sham, P. C. Model-free analysis and permutation tests for allelic association. *Hum. Hered.* **50**, 133–139 (1999).
51. Altshuler, D., Daly, M. & Kruglyak, L. Guilt by association. *Nature Genet.* **26**, 135–137 (2000).
52. Kwok, P.-Y. Genetic association by whole-genome analysis? *Science* **294**, 1669–1670 (2001).
53. Rothman, K. J. & Greenland, S. *Modern Epidemiology*, 2nd edn (Lippincott Raven, Philadelphia, 1998).
54. Marth, G. *et al.* Single-nucleotide polymorphisms in the public domain: how useful are they? *Nature Genet.* **27**, 371–372 (2001).
55. Sherry, S. T., Ward, M. & Sirotkin, K. Use of molecular variation in the NCBI dbSNP database. *Hum. Mutat.* **15**, 68–75 (2000).
56. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).
57. Brookes, A. J. *et al.* HGBASE: a database of SNPs and other variations in and around human genes. *Nucleic Acids Res*. **28**, 356–360 (2000).
58. Krawczak, M. & Cooper, D. N. The Human Gene Mutation Database. *Trends Genet.* **13**, 121–122 (1997).
59. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
60. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).

Acknowledgements

We thank members of the Myers laboratory and the Stanford Human Genome Center for their support. H.K.T., N.J.R. and R.M.M are supported by the National Institutes of Health and the Reynolds Foundation.

Online links

DATABASES

The following terms are linked online to:

CancerNet: http://www.cancer.gov
breast cancer

LocusLink:
http://www.ncbi.nlm.nih.gov/LocusLink
BRCA1 | *MECP2*

OMIM: http://www.ncbi.nlm.nih.gov/Omim
Crohn disease | Rett syndrome

FURTHER INFORMATION

Allele Frequency Project:
http://snp.cshl.org/afp_summary.html
Celera: http://www.celera.com
Genaissance:
http://www.genaissance.com/home_index.html
HGVbase (and SNP-related databases):
http://hgvbase.cgb.ki.se
Human Gene Mutation Database:
http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html
HGMD locus-specific mutation databases:
http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html
Access to this interactive links box is free online.