The Pharmacogenomics Journal (2002) 2, 349–360

www.nature.com/tpj

PERSPECTIVE

Interpretation of genetic association studies in complex disease

H Campbell[1] and I Rudan[1]

[1]*Department of Community Health Sciences, University of Edinburgh, Edinburgh EH8 9AG, UK*

The Pharmacogenomics Journal (2002) **2,** 349–360. doi:10.1038/sj.tpj.6500132

INTRODUCTION

The potential of the genetic association approach for the identification of genetic variants that alter susceptibility to common complex disease is well recognised.[1] This potential equally extends to the identification of genetic variants in genes coding drug-metabolising enzymes, transporters, receptors, and other drug targets that may determine inter-individual differences in drug responsiveness or the frequency of adverse drug reactions. Technological advances such as the availability of single nucleotide polymorphism (SNP) databases and affordable, very high throughput genotyping are set to extend the potential and improve the efficiency of association approaches. However, the very large number of genetic variants in the human genome[2] and the lack of detailed knowledge about the molecular and biochemical processes involved in aetiology of complex diseases or in drug response suggest that it is very likely that many spurious associations will be found and reported. The great majority of reported associations have not led to new insights into complex disease or drug response mechanisms. Important exceptions include genetic variants whose effects are large enough to be identified by linkage analysis (eg), variants in NOD2 in Crohn's disease,[3] APOE in Alzheimer's disease[4] and factor V Leiden in deep venous thrombosis[5]) and genetic variants in the cytochrome P450 CYP3A5 gene, which contribute to variation in bioavailability and clearance of drugs such as HIV protease inhibitors and some cholesterol-lowering drugs, and can result in drug toxicity.[6]

The primary aim of these studies is to identify significant associations between genetic variants and disease states, physiological or disease traits or markers of drug response or toxicity and then to judge whether these associations are 'causal' (truly alter susceptibility to disease or drug response). However, deciding which statistically significant associations are indeed causal is likely to represent a major obstacle to successful 'gene discovery', at least until the time when molecular pathways from gene to disease become better understood. In this review we discuss the correct use and limitations of existing criteria for the interpretation of association studies.

INTERPRETATION OF GENETIC ASSOCIATION STUDIES

In the most commonly employed (inductive) approach an assessment is first made of the validity of the observed association between the genetic variant and the disease or trait (*statistical inference)*. This involves considering the likelihood that alternative explanations of chance, bias and confounding could account for the findings. Secondly, all available biological and epidemiological evidence should be assessed to decide if the association is likely to be causal (*causal inference)*. This judgement is essential in order to decide what action is merited based on the results: guiding prioritisation of investment in future research, clinical management decisions or public health policy choices. This approach is one of *inductive* inference (which involves judging whether there is support for inferring a casual association based on the observed data).

STATISTICAL INFERENCE

Chance, bias and confounding are all alternative explanations for observed associations (Table 1).

Chance

A statistically significant result does not mean that chance cannot have accounted for the result, only that this is unlikely. It is a composite measure that reflects both the size of the difference and the sample size. Statistical significance testing does not give a yes/no answer but acts as a guide to whether the hypothesis or reported association is likely to be worthwhile pursuing further.[7]

Multiple testing is a major reason for published false positive reports. The expected frequency of false positives is given by $1-(1-k)^m$ (where m is the number of independent markers and k is usually <0.05, the significance level set for a single marker).[8] Reporting results of secondary and post hoc subgroup analyses as if they related to *a priori* hypotheses and selectively reporting only analyses that reach 'statistical significance' lead to covert multiple testing. New developments such as massive candidate gene analysis, genome screening by genetic association and adoption of pattern recognition methods such as artificial neural networks[9] will exacerbate the problem of multiple testing. Adopting statistical significance levels appropriate to the number of tests carried out can be used to limit the reporting of chance findings. In genetic linkage analysis of Mendelian traits, the adoption of the LOD score threshold of 3 or

Table 1 Examples of ways in which chance, bias and confounding can lead to false positive associations between genetic variants and disease states or traits or drug response

Chance
- Due to multiple association studies performed with publication of only those that show positive results (multiple testing together with publication bias)
- Due to testing of multiple markers (each with low prior probability of causing disease) and failure to adjust significance levels accordingly or otherwise interpret results appropriately
- Multiple testing due to reporting secondary and post hoc (subgroup) analyses as if they related to *a priori* hypotheses then selectively reporting only analyses that reach statistical significance

Bias
- Due to artefact (differences between cases/controls unrelated to cause of disease) such as differences in handling or storage between cases and controls
- Due to systematic error introduced in selection of cases and controls for study
- Due to systematic error introduced by differences in genotyping between cases and controls

Confounding
- Due to population stratification
- Due to differences in distribution of genetic and environmental risk factors for disease under study between cases and controls (limited, in theory, by ''Mendelian randomisation''[10])
- Due to linkage disequilibrium (LD) or 'allelic association' between marker under study and true disease susceptibility variant

more has been effective in reducing the number of false positive reports to below 5%.[11] Unfortunately, there is no similar international consensus for the interpretation of genetic association studies.[12–14] The Bonneferoni correction for multiple testing assumes that all variants being tested have equal prior probability and takes no account of the dependence that exists between adjacent variants. It thus leads to overcorrection, risking rejection of important findings. Schork has proposed a method to estimate the probability distribution of genetic association (case-control) test statistics empirically so that significance of genetic variants being studied can be assessed against this distribution. This method, however, is likely to be population and genome region specific.[15] An alternative strategy would involve grouping the variants to be tested into groups with differing prior probabilities (eg, variants with known effects on protein function would have higher prior probability) and applying empirical Bayes or semi-Bayes adjustments.[16,17] Greenland has noted that such a Bayesian analysis leads to a modest loss of power but a 'dramatic reduction in type 1 (false positive) error… by the use of prior information'.[18] As the functional significance of variants is better understood, so it will be increasingly possible to adopt an informed approach to adjustment of significance levels.[16] This principle is similar to the practice in clinical genetic risk counselling in which the significance of genetic variants is interpreted through the use of other relevant genetic information.[19]

Statistical tests of association are not strictly valid when there is dependence between individuals due to cryptic relatedness, which may or may not be apparent in recently collected pedigree data. This will lead to false positive associations particularly in inbred populations and in studies of rare disorders unless statistical methods that detect and account for these relationships are used.[20–22]

Bias

Any systematic differences in allele frequencies between cases and controls can result in an apparent association. Sources of bias (for example in the selection of study population or measurement of variables in cases and controls) have been discussed in detail.[23–25] In general, study designs that are prospective and in which case and control ascertainment is truly population based are more robust. The presence of artefacts leading to information bias can be explored by checking that genotype frequencies among controls are in Hardy–Weinberg equilibrium. Failure to find this draws attention to a problem with the selection, storage or analysis of control specimens and may suggest invalidation of the results of the association study.

Publication bias leads to the publication of relatively small initial studies selectively reporting large effects since they do not have adequate power to identify smaller effects. Studies reporting smaller effects do not reach statistical significance and so remain unpublished.[26] Subsequent more powerful and often better designed studies report more valid findings that either fail to reproduce the initial report (which then represent a false positive report due to a combination of chance and publication bias) or support the initial findings but with a more accurate (and modest) effect size.[27,28]

Confounding

Confounding factors are those that are associated with *both* the disease and the factor under study. Thus an apparent association between a genetic variant and a disease or drug reaction may be explained by confounding.[24] The size of the effect of a confounding factor is related *jointly* to its association with the factor under study and to the outcome. Multiple factors can contribute small amounts of confounding that together are substantial.[29] Strategies to control for confounding are limited since they can only be applied to factors that are currently known.

Most discussions of confounding in genetic association studies have focussed on *population stratification* although examples of its importance are few.[30] Biologically, plausible levels of population stratification are likely only to result in weak associations.[31,32] Approaches to limit or control population stratification include the use of family-based controls. However, the advantages of this approach have to

be weighed against the difficulty in recruiting parents (for diseases with onset in middle age) and the loss of power in comparison to case-control approaches. 'Genomic control' approaches utilise data from unlinked genetic markers to measure and adjust for population substructure effects.[21,33] These can be incorporated readily into studies of multiple candidate variants and are likely to be increasingly adopted.

Linkage disequilibrium (LD) between a (marker) variant under study and the true disease-susceptibility variant can result in confounding. High levels of LD within a population will increase the potential for confounding. Thus, although it has been proposed that initial association (LD mapping) studies be undertaken in small founder-pool populations with extended LD, confounding will complicate interpretation of findings. In addition to concerns about the irregular fine structure of LD,[2] frequent gene conversion disrupting regions of LD and the complex relationship between genetic and physical distance,[34,35] all positive associations will need to be followed by investigation of association in nearby variants and in surrounding haplotypes.[36] Fine mapping, for example by multiple candidate allele association analysis, will be more efficient in populations with low levels of LD, such as found in African populations.[37] High levels of LD (and thus confounding) within a population are likely to be regarded increasingly as a negative rather than a positive population attribute for genetic association studies.

CAUSAL INFERENCE

If chance, bias and confounding are all considered to be unlikely explanations for an observed association (Table 2) then it can be considered valid. A systematic approach to assessing whether valid associations may be *causal* can then be employed. Such approaches cannot *prove* causality since inference of cause from empirical data has no logical basis.[38] In addition, the aetiological heterogeneity and multifactorial nature of common complex disease, in which most factors under study will individually be neither necessary nor sufficient to cause the disease, complicates any approach to assessment of cause and effect relationships. Nevertheless, a set of criteria proposed by Bradford-Hill[39] based on the inductive canons of John Stuart Mill have proven useful (Table 3) and the utility of these are discussed below.

Consistency of Association

This approach parallels the successful strategy in linkage analysis whereby initial reports (with LOD score >3) need to be corroborated by an independent study to ensure a sound basis for genetic risk calculations.[40] In the absence of a broader understanding of genetic effects at the molecular level and of biochemical and physiological mechanisms, this criterion might seem the most powerful evidence in favour of causality currently available.

Replication of an association in the *same population* either in a 'split sample' or repeat independent sample gives evidence in favour of the variant being a causal variant. However, repeatability (probability that a second association study is also positive in the same population) varies with sample size and the proportion of trait variance attributable to the variant under study. Simulation has shown repeatability to be low with sample sizes of

Table 2 Appraisal of published associations with genetic variants: list of questions to consider in assessing validity of association

Chance
- Is it clear whether reported results relate to *a priori* hypotheses or post hoc subgroup analyses?
- Is the total number of analyses (number of tests) that were carried out stated?
- Has an adjustment of the statistical significance level to account for multiple tests (eg), Bonneferoni or Bayes methods) been made or has interpretation of results otherwise accounted for multiple testing?
- Does statistical analysis account for increased likelihood of chance association in inbred populations or, where relevant, due to cryptic relatedness in apparently outbred populations?
- Where no statistically significant association was found, was the sample size large enough for adequate (eg, 80%) power to detect important/plausible effect sizes?

Bias
- Were the genotype frequencies reported in the control specimens in Hardy–Weinberg equilibrium? If this was not the case, were the reasons for this explored? Could this signal the presence of bias or study artefacts?
- Are the procedures for the ascertainment of cases and controls carefully described; could they have resulted in bias that could explain the results?
- Is the control group drawn from the same population as the cases?
- If 'convenience' controls were used (such as blood donors) is information presented on the degree to which they are representative of the population from which the cases are drawn? Could these differences explain the results?
- If published control allele frequencies were used to give control data, was their appropriateness in this study population reviewed critically? Would adoption of alternative allele frequencies alter the results?
- Are participation rates in cases and controls stated? If substantially different could this explain the results?
- Are there sufficient details of the study procedures (handling and storage of blood and DNA specimens or analysis; genotyping methods; other measurement methods) for both case and control specimens? Were the methods valid and consistently applied? Were there any systematic differences in procedures between cases and controls? Could any differences have accounted for the results?

Confounding
- Were attempts made to limit any effects of confounding factors such as population stratification by
  - Restriction of the study population (eg, use of family-based control approaches)
  - Matching on reported ethnicity or adjustment for factors in a stratified or multivariate analysis (eg, genomic control methods)

Table 3 Definition and utility of criteria for identification of 'causal' associations between genetic variants and a disease state or trait

Criterion	*Definition*	*Utility*
Consistency of association	• Consistent association with a genetic variant across studies in the *same population* provides evidence in support of causal association • Replication of association in *different populations* is based on assumption that the same genetic variant will be a causal factor and will be detectable in other populations and that other required (genetic or environmental) component causes will also be present • One good study (well designed in terms of limiting bias and confounding and of large sample size) outweighs several poor ones • Replication in studies with different study designs provides stronger evidence of causality (thus, consistent evidence both from association studies with population- and family-based controls improves evidence for causality)	Studies on independent data from the same population • Replication good evidence of causal association • Lack of replication against causal association only if sample size is adequate (see discussion of repeatability in text) Studies in different populations • Replication good evidence of causal association • Lack of replication difficult to interpret as different genetic background, environmental exposure and LD patterns may lead to different causal genetic variant in other populations (do not reject initial findings on this basis alone) Studies in special populations • Replication of findings from special populations problematic (risking important findings being rejected)
Strength of association	• Strong association is better evidence for a causal relationship than a weak one as it is less likely that bias and confounding can explain a strong effect • Strength of association depends on the particular study population and the prevalence of other causal factors (genetic and environmental)	• A strong association is better evidence for a causal relationship than a weak one although utility limited by over-estimation of effect size especially in initial reports • Strong association may be reported for a marker in strong LD with susceptibility variant of weak effect and vice versa • Strong association in family-based studies does not imply tight linkage • Strong association does not equate to importance as a cause of disease in the population as population attributable fraction depends also on the allele frequency of the variant
Biological plausibility	• If a relationship is consistent with knowledge of mechanisms of the disease then it is more likely to be causal. • Conversely, if there is no known biologically plausible mechanism but epidemiological evidence is otherwise strong then this probably reflects limitations of medical knowledge	• Association generally consistent with known molecular mechanism of the disease currently represents weak support for causality as many plausible mechanisms can be constructed post hoc • Formulation of objective and quantifiable criteria should provide more robust evidence favouring casuality; thus, future adoption of rigorous criteria from bioinformatics (biological sequence comparison and computational gene and protein structure prediction) will lead to more critical and evidence) based application • Biological data can be used to direct genetic analysis, eg, defining criteria for rational selection of candidate genes for study thus reducing multiple testing and focusing on variants with higher prior probability of pathological role
Biological gradient (or dose–response relationship)	• Varying amounts of a factor related to varying amounts of effect (eg, greater risk of disease, earlier disease onset or more severe disease) with an observable gradient strengthens evidence for causal association • Good evidence of causality if the association is strong, but does not exclude confounding when association is weak	• Presence dependent on underlying genetic model; may be threshold effect with multiple susceptibility variants interacting to disturb homeostatic mechanisms • When present can give useful information about the genetic model that is operating
Temporal relationship	• Cause to precede effect	• Not generally helpful in assessing cause • May have some utility in instances in which epigenetic mechanisms such as methylation and adduct formation are known to be present
Analogy	• Existence of well-known cause analogous to one under study	• Known cause analogous to one under study not helpful in assigning cause
Reversibility	• Removal of factor results in decreased disease risk	• Not generally applicable

Table 3 (Continued)

Criterion	*Definition*	*Utility*
specificity	• A single cause leads to a single effect	• Very unlikely to be valid for most genetic variants underlying complex disease since most show substantial pleiotropy (more than one phenotype determined by same genotype) • Known to hold only for infectious diseases and some inborn metabolic errors
Experimental evidence	• Experimental evidence is best considered as a test of a hypothesis of causal association	• Supportive data from functional studies such as knock-out animals, cell lines, and studies of gene expression and enzyme activity are strong evidence in favour of causality • Data showing genetic variant under study is expressed in diseased tissue or alters enzyme or receptor activity in relevant metabolic pathway strengthens evidence in favour of causality

100–500 cases[41] and empirical data confirm this with studies of sample size less than 150 followed much more often by studies reporting discrepant results.[27]

Seeking replication of an association in *another population* is only valid if it is likely that the two data sets share the same measurement value (within sampling variability).[40] Substantial between-study heterogeneity has been clearly shown in a review of repeated genetic association studies.[27] Thus, the prospects for replication are uncertain when the validation sample differs genetically and/or environmentally from the original study population.[42] Different proband ascertainment strategies, multiple disease alleles, outbreeding and environmental modifiers all act to make replication of findings less likely.[40] Where LD is the basis of the observed association, this is not likely to be consistent across populations since LD depends on population history. Even when the 'causal' disease susceptibility variant is under investigation, a genetic variant may be more or less important in different populations depending, for example, on population allele frequencies. It may prove particularly problematic to replicate associations reported in 'special' populations (genetic isolate, admixed or those with unusual environmental exposure patterns) especially if the variant studied has low relative risk, variable penetrance and very variable allele frequencies in different populations. More generally, rare variants (<5% population prevalence), which may be particularly important in the aetiology of complex disease,[43,44] are more likely to be population-specific. Replication in another population of associations with rare variants may not be possible.[45] In these circumstances other alternative (functional) variants in the gene under investigation should be studied. Positive associations with these variants could represent evidence in favour of a causal role. This is similar to finding family specific mutations in linkage studies of a Mendelian disease. Risch has suggested that such allelic heterogeneity provides strong evidence of a causal relationship.[10]

Replication studies must therefore ensure they have a sufficiently large sample size to give adequate power to detect the association (see also tendency to over-estimate effect sizes and hence study power[42] below). Due to the problems with replication in other populations, integrated study designs[46] that permit an internal check in an independent sample of the same population should be favoured. This could include, for example, designs that include cases and both population- and family-based controls. Repetition by a transmission disequilibrium test (TDT) study following a reported association in a case-control study demonstrates both linkage and association and would further strengthen the evidence for a causal association.

When an association is confirmed in other populations then chance is a highly unlikely explanation. However, failure to confirm the association is more problematic to interpret, as discussed above. Rejection of findings not replicated in other populations may discard genetic effects with important effects specific to population subgroups.

Strength of Association

In complex causal pathways with multiple interacting causes (none of which might be either necessary or sufficient), associations tend to be of modest strength and inferences based on the relative strength of individual estimates of relative risk are problematic. Added to this, the application of this criterion in judging whether an association may be causal is complicated by a number of factors which bias reports of association strength. A consistent upward bias in published estimates of locus-specific effect sizes has been noted. This is due to publication bias in initial reports[27] and also due to the 'Beavis effect', particularly when maximum likelihood methods are employed.[42] In LD mapping studies the strength of association is influenced by the extent of LD between and the relative frequencies of marker and susceptibility variants. Thus, strong associations may be reported for a marker in strong LD with

susceptibility variant of weak effect and vice versa.

It should be noted that in family-based associations studies results are based on recombinations occurring in a single generation and so cannot distinguish between tight and loose linkage. Thus strong association in these studies does not imply tight linkage.[47]

Biological Plausibility

The molecular nature of the genetic variant may guide interpretation of an observed association with a disease, disease trait or adverse drug reaction (Table 4). Where appropriate, the adoption of an underlying biological model (for example, the multistep model of carcinogenesis) may provide a useful framework for interpretation. As the function of specific genes and their role in biological processes become better understood, it will be increasingly possible to direct 'candidate gene' studies based on this knowledge. This Bayesian approach would favour investment in (persisting with) investigation of variants in which there are prior biological reasons to suspect a role for a candidate gene (Table 4). For example, genes that are more highly expressed (high number of mRNA copies) in tissues in which disease pathology is known to occur could be selected first for study.[48] This has been shown to result in a 30 to 100-fold reduction in the number of genes to be screened.[49]

It is likely to be more efficient to investigate SNPs in coding and promoter regions or SNPs that define 'haplotype tags'[50] than random SNPs. Typologies of SNPs similar to the classification system in Table 4 have been developed.[10,51,52] Critical biochemical processes that are well defined and under the control of both genes and environmental exposures might prove to be good starting points for the investigation of the role of genetic factors in pathogenetic mechanisms.

Current approaches to the assessment of biological plausibility are subjective and unsatisfactory. They are typically based on prior beliefs or involve post hoc biological hypotheses being drawn up by investigators keen to find support for an observed association. Emerging bioinformatics methods in biological sequence comparison, computational gene prediction, identification of functional gene signals and prediction of protein structure[53] should allow Bayesian methods to quantify 'biological plausibility' on a probability scale. This will permit biological plausibility to be assessed objectively in a scientific manner and will greatly improve the utility of this criterion in causal inference.

Biologic Gradient (Dose–Response Relationship)

The presence of a gradient supports the interpretation that the variant truly alters susceptibility to disease, although association due to confounding factors can also show a gradient. However, the presence of a gradient is dependent on the underlying genetic model. Thus, if there is a moderate risk of disease in heterozygotes and (very) high risk in homozygotes, this not only favours causal association but provides information on the underlying model (in this case recessive). Conversely, interpretation of results is complicated by lack of knowledge of the particular underlying genetic model that is operating and so lack of a gradient is not necessarily evidence against causal association. A threshold effect may be seen in which no effect is observed until there is a certain level of 'exposure' — with genetic factors this may be through multiple

Table 4 Examples of the synthesis of epidemiological and biological data in the design or interpretation of genetic association studies

Study Design

Use of both epidemiological and biological data to define criteria for selection of candidate genes

Strong justification:

- Genetic variant associated with familial forms of disease
- Genetic variant in exon or intronic promoter region of gene coding for proteins involved in molecular mechanisms of disease or for xenobiotic enzymes thought to interact with environmental exposures known to mediate risk of disease
- High mRNA copy number in tissues affected by pathological process

Weak justification:

- Genetic variant found to be associated with disease risk in other published reports but no other supporting biological data

Use of data from genetic association studies to direct functional investigation of candidate genes ('statistical functional genomics')[57]

- Enumeration of all genetic variants in a genomic region
- Results of genetic epidemiological analysis direct and prioritise subsequent molecular and functional studies

Data analysis/interpretation

Use of biological criteria to classify genetic variants into categories with differing probabilities of having a true pathological role:

Probability of pathological role	Variant (type of mutational event)
• Definitely pathogenic	• Frameshift; nonsense; splice variant
• Probably pathogenic	• Nonconservative amino acid change
• Probable polymorphism	• Conservative change; variant in controls
• Definite polymorphism	• Synonymous variant

susceptibility variants interacting to disturb homeostatic mechanisms and thus alter a trait value.

Temporal Relationship

This criterion is apparently self-evident in genetic studies since the genotype is fixed from conception and thus always precedes disease or drug response. However, consideration of this criterion may be relevant in the study of epigenetic changes such as methylation of DNA or DNA adducts (from carcinogens, which can occur in response to environmental exposures later in life) or in gene expression studies (with genes up and down regulated). Evidence that specific epigenetic changes occurred before the earliest pathological changes would be consistent with an interpretation that they may have caused the changes.

Specificity

Genetic variants are likely to have pleiotropic effects and thus would not be expected to show highly specific pathological effects. However, this criterion may be helpful in a broader sense, for example, if there is a large pedigree or an unusual (isolate or admixed) population in which a specific subtype of a complex disease is found. This may be represented as an extreme incidence or prevalence of disease or as very early onset disease or very severe disease. If related pathophysiological and biochemical evidence confirms a specific subtype of the complex disease then population- or pedigree-specific variants may lead to very specific forms of disease and this may provide evidence in favour of causality.

Analogy

This is the weakest criterion, as analogies can be readily found everywhere.

Table 5 Interpretation of genetic association studies by deduction: use of deductive criteria to judge among competing hypotheses to explain observed association

		Competing hypotheses for interpretation of data				
Criteria (based on observed data)		*False positive (chance)*	*Artefact (bias)*	*Population stratification (confounding)*	*LD with causal variant (confounding)*	*Causal variant*
Consistency[a]						
Replication in same population	Present	**Against**	Against[b]	Against[c]	In favour	**In favour**
	Absent	**In favour**	In favour	In favour	Against	**Against**
Replication in different population	Present	**Against**	**Against**	**Against**	Against	**In favour**
	Absent	—	—	—	In favour	**—**
Replication in different population with other variant at same locus	Present	Against	Against	Against	Against	**In favour**
	Absent	—	—	—	In favour	**—**
Strong association[d]	Present	Against	Against	**Against**	Against	**In favour**
	Absent	—	—	—	—	—
Biological plausibility[e]	Present	—	—	—	—	—
	Absent	—	—	—	—	—
Biological gradient	Present	—	—	—	Against	In favour
	Absent	—	—	—	In favour	—
Control alleles in Hardy-Weinberg equilibrium	Present	—	Against	Against	—	—
	Absent	—	**In favour**	**In favour**	—	—
Association persists after appropriate correction made for any multiple testing	Present	**Against**	—	—	—	In favour
	Absent	**In favour**	—	—	—	Against
Test for association in surrounding variants	Present	**Against**	In favour	In favour	In favour	Against
	Absent	—	Against	Against	Against	In favour
Experimental data (functional studies)	Present	**Against**	**Against**	**Against**	**Against**	**In favour**
	Absent	**—**	**—**	**—**	**—**	**Against**

Test of causal hypotheses. A list of competing explanations (or hypotheses) for the association is set out. These can be found in the columns (eg, chance, population stratification, linkage disequilibrium with causal variant or causal association with the variant showing association). It is assumed by deduction that one of these is correct. The competing explanations are then tested against the observed data by considering the criteria listed in the rows. If only one remains unrefuted then it is considered to be correct. The comments in the table against/in favour represent evidence against/in favour of this interpretation; and **against/in favour** represent **strong evidence** against/in favour of this interpretation. This is similar to an outbreak investigation in that possible exposures that may have caused the outbreak are eliminated one by one (eliminative induction).[38]

[a]Assuming that there is a sufficient sample size to assure good probability of repeatability of true finding.

[b]Evidence against this interpretation if replication study used different study design (for example family-based controls in repeat study if original used population-based controls).

[c]Evidence against this interpretation if replication study used family-based controls.

[d]Definition of 'strong' association in a complex disease will vary: an unbiased odds ratio of greater than 2 is strongly against an interpretation of population stratification; an unbiased odds ratio of greater than 3–5 is strongly in favour of a causal variant.

[e]Assuming subjective assessment of biological plausibility; evidence in favour/against is stronger if objective quantified assessment of biological plausibility is made.

EXPERIMENTAL EVIDENCE

Experimental evidence is best regarded as a *test of a causal hypothesis* (see Table 5) than as a criterion for causal inference. In the early genetic linkage analysis studies, results were corroborated by cytogenetic or somatic cell hybrid studies in order to yield robust conclusions that could be used to direct clinical genetics risk estimations. The limitations of the utility of epidemiological evidence alone in determining causality are illustrated in the above discussion. This highlights the need to integrate epidemiological and statistical data with biological data in order to build a more robust framework for interpretation. A general framework for synthesising epidemiological and experimental data is illustrated in Figure 1. Genetic variants showing positive associations with disease or disease traits that appear to be causal should be examined further in functional studies in knock-out animals or cell lines or in gene expression or enzyme activity studies, as appropriate. The discovery of the role of the mismatch repair genes in a subset of early onset colorectal cancer followed this model (Figure 2).

ALTERNATIVE APPROACHES TO USE OF ABOVE CRITERIA IN INTERPRETATION OF GENETIC ASSOCIATION STUDIES

The traditional epidemiological approach described above is one of assessing the validity of the association (by considering chance, bias and confounding;Table 2) and then applying *inductive inference*. This assesses the extent to which the data support an interpretation (or hypothesis) that the exposure or genetic variant under study is a cause of the disease (Table 3). An objective assessment is made of each of the above criteria leading to a judgement as to whether the weight of the evidence is in favour with this interpretation. This approach, however, is greatly compromised by the problems in interpretation highlighted above.

An alternative and preferred approach is as a *deductive* test of causal hypotheses. A list of competing explanations (or hypotheses) for the association is set out. It is assumed by deduction that one of these is correct. The competing explanations are then tested against the observed data. If only one remains unrefuted then it is considered to be correct. This is similar to an outbreak investigation in which possible exposures that may have caused the outbreak are eliminated one by one (eliminative induction).[38] An illustration of this approach is given in Table 5. This deductive approach has a more secure basis in logic.[38] It can also help identify what kind of data or further analysis may be useful to distinguish between competing explanations (hypotheses) and thus help direct further research. The design of biological experiments to test key causal hypotheses has typically

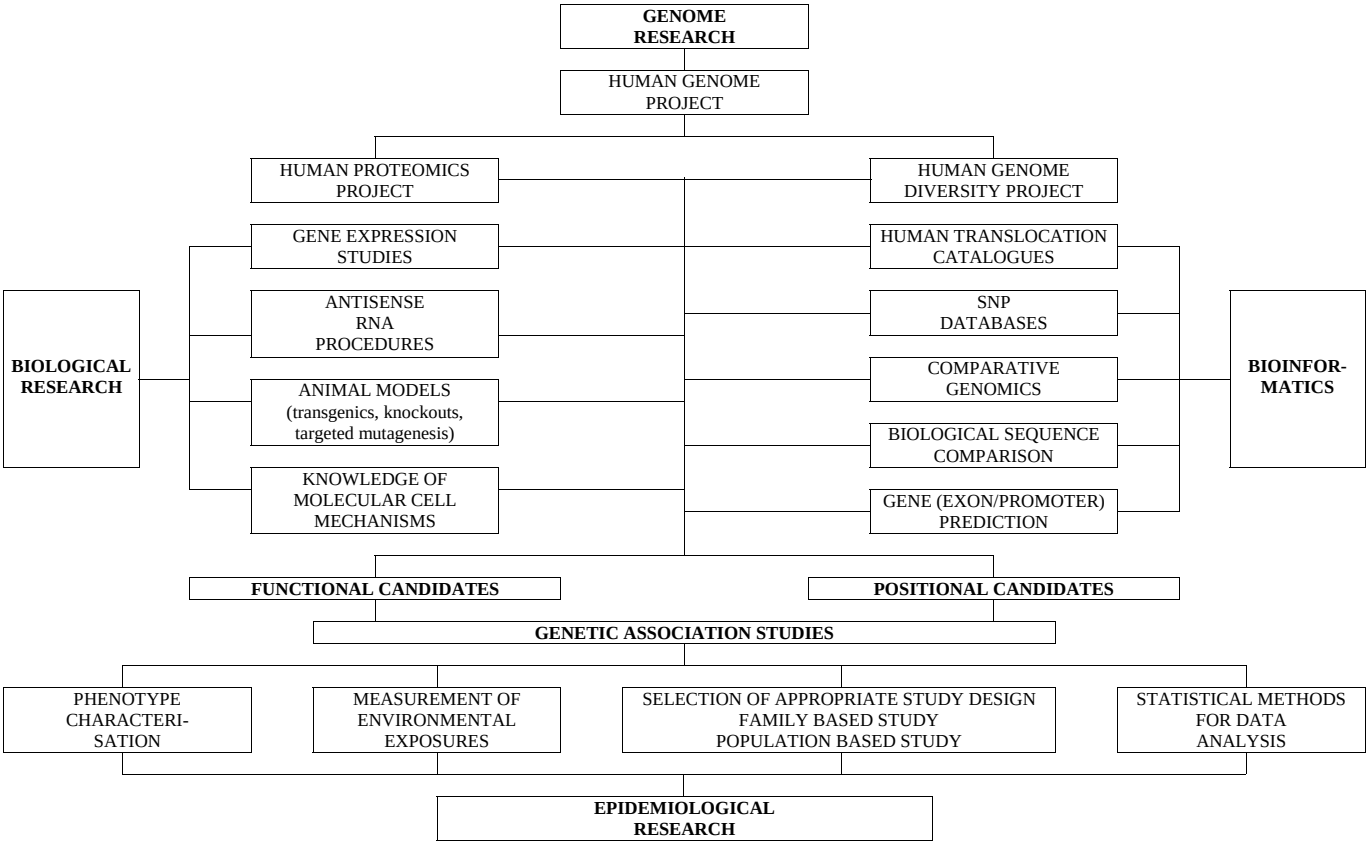

Figure 1 Synthesis of epidemiological, genomic, biological and bioinformatics data in the design of genetic epidemiological studies.

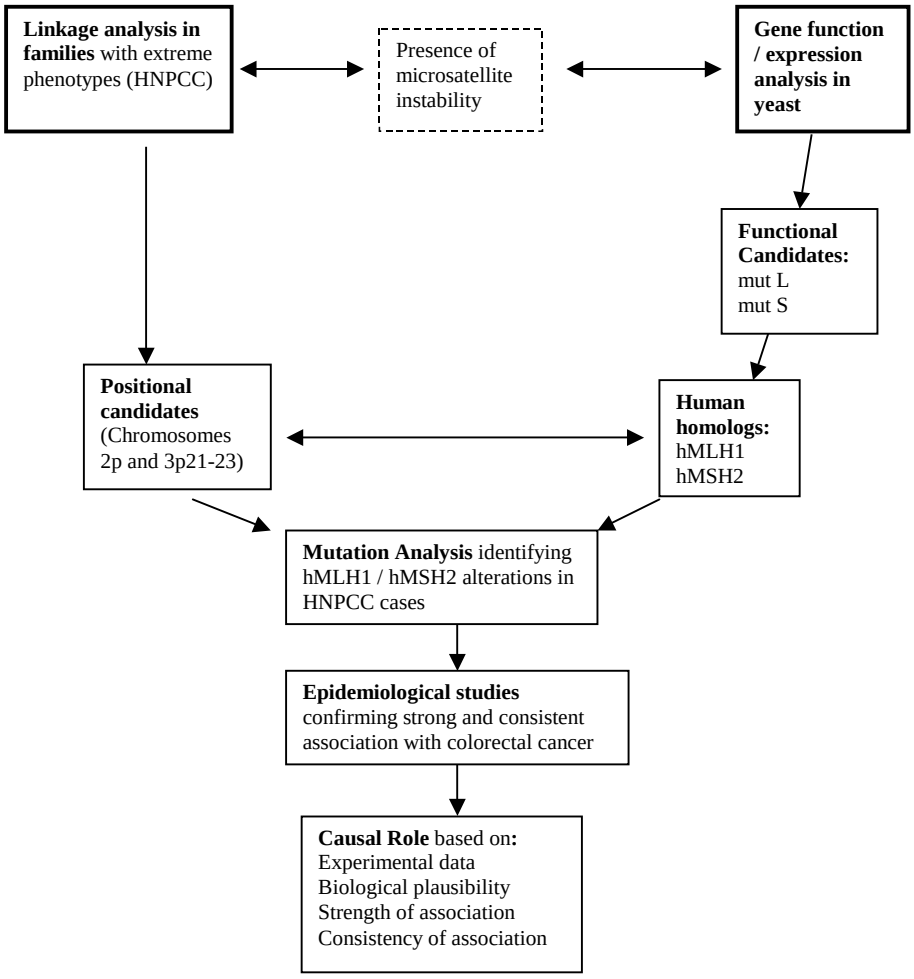

Figure 2 Synthesis of epidemiological and biological data: discovery of the role of human mismatch repair gene mutations (hMLH1 and hMSH2) in colorectal cancer.

not been feasible in past epidemiological studies. This has placed severe limits on the approach to interpretation based on inference. The design of experimental studies (such as functional studies of genetic variants) is now increasingly possible in genetic epidemiology (Figure 2). This should be seen as an *essential step* in the interpretation of genetic association studies and should ideally be planned by multidisciplinary teams including epidemiologists and statisticians together with geneticists and biologists. The expense of these studies is likely to be offset by the savings resulting from fewer research groups pursuing false positive associations.

CONCLUSIONS

The failure of the majority of reported genetic associations to lead to new insights into complex disease or drug response mechanisms challenges the perceived utility of this approach for the identification of genetic variants underlying common complex disease or responsiveness or adverse reaction to drugs. Closer attention to study design and use of bioinformatics data to inform data analysis (through empirical Bayes adjustments) has the potential to limit the rate of false positive reports. Suggestions that comparisons of groups of individuals defined by genotype in a genetic association study are equivalent to randomised comparison (due to ''Mendelian randomisation'') and thus not be susceptible to bias and most confounding effects are potentially important but need to be demonstrated empirically.[10] If the results of genetic association studies are to provide a useful guide to direct further research then an appropriate framework is required to assess whether genetic variants showing an (apparently valid) association with disease truly alter susceptibility to disease or response to drugs. Ideally, knowledge of genetic variability should inform study design and interpretation and this is likely to evolve as the Human Genome Diversity Project matures. Bioinformatics strategies should improve our ability to utilise biological knowledge to

Table 6 Linked epidemiological and bioinformatics approaches for correct interpretation of genetic association studies

	Epidemiology	*Biology/Bioinformatics*
Study Design	Characterise phenotype and measure relevant environmental exposures accurately and precisely	Select variants that are most likely to have functional effects in important biological pathways for investigation
	Select appropriate study population to answer study hypothesis (family-based study — multiplex family or affected family members; population-based study; special population — isolate or admixed; twins)	Select variants identified as positional or functional candidates from prior biological research (see Figure 1) for investigation
	Adopt appropriate study design and procedures to limit bias and confounding	
	Recruit sufficiently large sample to ensure adequate power to detect modest association with genetic variant and to permit replication of finding within an independent subset of the data	
Data analysis and interpretation of positive associations	Identify all adjacent genetic variants and check for association with disease *(if no association then initial association more likely to be causal)*	Quantify probability of variant having relevant functional effects through formal bioinformatics procedures such as biological sequence comparison and gene and protein prediction programs *(to provide objective and quantified assessment of biological plausibility)*
	Identify all adjacent genetic variants and check for association among variants in control chromosomes *(if significant allelic association, comparison of haplotype frequencies may be better)*	Assess functional consequences of genetic variants showing association with disease/trait under study
	Adopt appropriate statistical significance levels or otherwise interpret findings according to the number of tests performed	Investigate potential to check association through experimental studies in animal models *(such as transgenics or knockouts to look for confirmatory evidence of functional effect)*
	Check whether control genotype distribution is in Hardy–Weinberg equilibrium to check for study artefacts or biases.	Look for confirmatory evidence from gene expression studies *(high levels of gene expression in tissues known to be affected by disease supports role in disease susceptibility)*
	Seek to replicate the finding in the same or different study population	
	Quantify the size of effect associated with the genetic variant and look for evidence of biological gradient in effect	

direct epidemiological studies and inform interpretation of results. This will build a more scientific, evidence-based approach to the consideration of *biological plausibility* than the current unstructured and unhelpful approach and will result in this criterion becoming more useful in future. *Consistency of association* across studies is a useful indicator of causal association, when present. However, problems in interpreting failure to replicate findings limits the utility of this criterion and argues in favour of investment in large, integrated study designs that can perform internal checks in a single population. The difficulties in replicating associations in special populations or with rare variants should be recognised if important population-specific effects (which may give unique insights into molecular processes relevant to disease in all populations) are not to be discarded. Furthermore, meta-analyses to determine a summary measure of association across different populations may underestimate the effect of variants in specific populations. *Strength of association* remains a useful indicator of causal association but over-estimates due to bias are frequent. A deductive framework based on testing competing

causal hypotheses and involving both epidemiological and experimental data (for example from animal models or from gene expression or functional studies) is proposed.

This review seeks to highlight the need for improved strategies for the interpretation of genetic association studies. In particular, the development and support of multidisciplinary groups with expertise in bioinformatics, (genetic) epidemiology/statistics and experimental biology is important for the future success of this field. Moving towards this model in which epidemiological and experimental biological data are synthesised together (Tables 4–6, Figure 1) will require recognition in the policies of research funding agencies and research funding to be redirected.[54]

Current proposals to move away from a hypothesis testing paradigm of investigation to one of high-throughput (functional) genomics[55,56] in which very large numbers of variants are related to a wide range of phenotypes underlines the need for an international consensus on a framework for the interpretation of genetic association studies and for the issues raised in this review to inform the design of these studies.

ACKNOWLEDGEMENTS

We would like to acknowledge helpful comments made by Dr Andrew Carothers and Professor Alan Wright, MRC Human Genetics Unit, Edinburgh; and the help of Rory Mitchell in preparing Figure 2. IR is supported by the UK Medical Research Council and grant funding from the Wellcome Trust.

DUALITY OF INTEREST

None declared.

Correspondence should be sent to
H Campbell, Department of Community Health Sciences, University of Edinburgh, Edinburgh EH8 9AG, UK.
Tel: +131 650 3218
Fax: +131 650 6909
E-mail: harry.campbell@ed.ac.uk

REFERENCES

1 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Glatt CE, DeYoung JA, Delgado S, Service SK, Giacomini KM, Edwards RH *et al*. Screening a large reference samples to identify very low frequency variants: comparisons between two genes. *Nat Genet* 2001; **27**: 435–438.
3 Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R *et al*. A frameshift mutation in Nod2 associated with susceptibility to Crohn's disease. *Nature* 2001; **357**: 1925–1928.
4 Kehoe P, Wavrant-De Vrieze F, Crook R, Wu WS, Holmans P, Fenton I *et al*. A full genome scan for late onset Alzheimer's disease. *Hum Mol Genet* 1999; **8**: 237–245.
5 Zoller B, Dahlback B. Linkage between inherited resistance to activated protein C and factor V gene mutation in venous thrombosis. *Lancet* 1994; **343**: 1536–1538.
6 Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Shuetz J *et al*. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genet* 2001; **27**: 383–391.
7 Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotech* 1998; **9**: 578–594.
8 Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; **2**: 91–99.
9 Lucek P, Hanke J, Reich J, Solla SA, Ott J. Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum Heredity* 1998; **48**: 275–284.
10 Youngman L, Keavney B, Palmer A, *et al*. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and 6002 controls: test of caisality by ''Mendelian randomisation''. *Circulation* 2000; **102** (Suppl. II): 31–2.
11 Risch N. Searching for genetic determinants in the new millennium. *Nature* 2000; **405**: 847–856.
12 Elston RC. Algorithms and inferences: the challenges of multifactorial diseases. *Am J Hum Genet* 1997; **60**: 255–263.
13 Kruglyak L. What is significant in whole-genome linkage disequilibrium studies?. *Am J Hum Genet* 1997; **61**: 810–812.
14 Morton NE. Significance levels in complex inheritance. *Am J Hum Genet* 1998; **62**: 690–697.
15 Schork NJ. Power calculations for genetic association studies using estimated probability distributions. *Am J Hum Genet* 2002; **70**: 1480–1489.
16 Greenland S, Robins JM. Empirical Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991; **2**: 244–251.
17 Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000; **29**: 158–167.
18 Greenland S. Probability logic and probabilistic induction. *Epidemiology* 1998; **9**: 322–332.
19 Petersen GM, Parmigiani G, Thomas D. Missense mutations in disease genes: a Bayesian approach to evaluate causality. *Am J Hum Genet* 1998; **62**: 1516–1524.
20 McPeek MS, Sun I. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* 2000; **66**: 1076–1094.
21 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
22 Newman DL, Abney M, McPeek MS, Ober C, Cox NJ. The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet* 2001; **69**: 1146–1148.
23 Miettinen OS. *Theoretical Epidemiology*. John Wiley: New York, NY, 1985.
24 Vineus P, McMichael AJ. Bias and confounding in molecular epidemiological studies: special considerations. *Carcinogenesis* 1998; **19**: 2063–2067.
25 Rothman K. *Modern Epidemiology*. Little, Brown and Co.: Boston, MA, 1986.
26 Dickersin K, Min Y-I, Meinhert CL. Factors influencing publication of research results: follow up of applications submitted to two institutional review boards. *JAMA* 1992; **267**: 374–378.
27 Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29**: 306–309.
28 Keavney B, McKenzie C, Parish S, Palmer A, Clark S, Youngman L *et al*. Large-scale test of hypothesised associations between the angiotensin-converting enzyme insertion/deletion polymorphism and myocardial infarction in about 500 cases and 6000 controls. *Lancet* 2000; **355**: 434–441.
29 Thomson WD. Statistical analysis of case-control studies. *Epidem Rev* 1994; **16**: 33–50.
30 Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci* 1998; **95**: 11389–11393.
31 Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000; **92**: 1151–1158.
32 Wacholder S, Rothman N, Caporaso NE, Garcia-Closas M, Buetow K, Fraumeni JF. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta* 2001; **1471**: C1–C10.
33 Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
34 Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J *et al*. Haplotype structure and population genetic inferences from nucleotide sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998; **63**: 595–612.
35 Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
36 Keavney B, McKenzie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E *et al*. Measured haplotype analysis of the angiotensin-1 converting enzyme. *Hum Mol Genet* 1998; **7**: 1745–1751.
37 Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ *et al*. Linkage disequilibrium

in the human genome. *Nature* 2001; **411**: 199–204.
38 Greenland S. Induction versus Popper: substance versus semantics. *Int J Epidemiol* 1998; **27**: 543–548.
39 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965; **58**: 295–300.
40 Vieland VJ. The replication requirement. *Nat Genet* 2001; **29**: 244–245.
41 Long AD, Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9**: 720–731.
42 Goring HHH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001; **69**: 1357–1369.
43 Wright AF, Hastie N. Complex genetic diseases: controversy over the Croesus code. *Genome Biol* 2001; **2**: 2007.1–2007.8.
44 Pritchard JF. Are rare variants responsible for susceptibility to complex disease? *Am J Hum Genet* 2001; **69**: 124–137.
45 Szabo CI, King MC. Population genetics of BRCA1 and BRCA2. *Am J Hum Genet* 1997; **60**: 1013–1020.
46 Zhao LP, Aragaki C, Hsu L, Potter J, Elston R, Malone KE *et al*. Integrated designs for gene discovery and characterisation. *J Natl Cancer Inst Monogr* 1999; **26**: 71–80.
47 Whittaker JC, Denham MC, Morris AP. The problems of using the transmission/disequilibrium test to infer tight linkage. *Am J Hum Genet* 2000; **67**: 523–526.
48 Cepko CL, Austin CP, Yang X, Alexiades M, Ezzeddine D. Cell fate determination in the vertebrate retina. *Proc Natl Acad Sci* 1996; **93**: 589–595.
49 Wright AF, Van Heyningan V. Short cut to disease genes. *Nature* 2001; **414**: 705–706.
50 Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, DiGenova G *et al*. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
51 Cargill M, Altschuler D, Ireland J, Sklar P, Ardlie K, Patil N *et al*. Characterisation of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999; **22**: 231–238.
52 Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A *et al*. Patterns of single-nucleotide polymorphisms in candidate genes for blood pressure homeostasis. *Nat Genet* 1999; **22**: 239–247.
53 Balding DJ, Bishop M, Cannings C. *Handbook of Statistical Genetics*. Wiley: Chichester, England, 2001.
54 Gambaro G, Anglani F, D'Angelo A. Association studies of genetic polymorphisms and complex disease. *Lancet* 2000; **355**: 308–311.
55 Goodman L. Hypothesis-limited research. *Genome Res* 1999; **9**: 673–674.
56 Yaspo M. Taking a functional genomics approach in molecular medicine. *Trends Mol Med* 2001; **7**: 494–501.
57 Almasy L, Blangero J. Challenges for genetic analysis in the 21st century: localising and characterising genes for common complex diseases and their quantitative risk factors. *GeneScreen* 2000; **1**: 113–116.