Candidate gene case-control association studies: advantages and potential pitfalls

Ann K. Daly[1,2] & Christopher P. Day[2]

[1]*Department of Pharmacological Sciences and* [2]*Centre for Liver Research, University of Newcastle upon Tyne, Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH*

There is increasing information on the importance of genetic polymorphisms in human genes. Polymorphisms occur on average once every 500–1000 base pairs in the human genome and are useful in the identification of genes involved in human disease. Some genetic polymorphisms have functionally significant effects on the gene product and are the most useful type of polymorphism in disease association studies while others are simply useful markers. There are two main approaches using polymorphisms in the identification of genes involved in polygenic diseases. The first involves examining inheritance patterns for genetic polymorphisms in family studies and the second case-control studies which compare genotype frequencies for candidate disease genes in unrelated individuals with the disease and healthy controls. Use of family studies is generally the preferred approach but this is only feasible if the genetic component of the disease is relatively strong, DNA samples are available from other family members and the disease is relatively easy to diagnose and is not stigmatized. Population case-control studies are useful both as an alternative and an adjunct to family studies. When performing case-control studies factors such as study design, methods for recruitment of cases and controls, functional significance of polymorphisms chosen for study and statistical analysis of data require close attention to ensure that only genuine associations are detected. To illustrate some potential problems in the design and interpretation of association studies, some specific examples of association studies on drug response and on disease susceptibility involving receptor genes, cytochrome P450 and other xenobiotic metabolizing enzyme genes and immune system genes including TNF-α, IL-10 and the IL-4 receptor are discussed.

Keywords: case-control study, cytochrome P450, cytokine, drug response, polymorphism

Introduction

Genetic polymorphisms occur throughout the human genome on average once every 500–1000 base pairs (bp). The normal definition of a genetic polymorphism is a variation in DNA sequence that occurs at least once in every 100 copies [1]. Except in the case of genes on sex chromosomes, each individual will have two copies of every gene so the variation needs to occur at least once in every 50 individuals to be classed as a polymorphism. Rarer sequence variation will also occur but these rare variants are usually referred to as isolated mutations rather than genetic polymorphisms. The normal DNA sequence for a particular gene is usually referred to as the wild-type allele and the rarer sequence referred to as a variant allele with the term allele defined as an alternative form of a gene. It is not uncommon for several different variant alleles to occur for a single gene. Different polymorphisms that occur close together are often linked, that is they are found on the same chromosome more frequently than would be expected by chance. The arrangement of genetic polymorphisms within a single chromosome in an individual is known as a haplotype. Frequently certain haplotypes are more common as a result of linkage

Correspondence: Dr A. K. Daly, Department of Pharmacological Sciences, University of Newcastle upon Tyne, Medical School, Framlington Place, Newcastle upon Tyne NE2 4HH. Tel.: 0191 2227031; Fax: 0191 2227230; E-mail: A.K.Daly@newcastle.ac.uk

Received 6 September 2000, accepted 1 August 2001.

than would be expected if each polymorphism was inherited randomly. In this case, the polymorphisms are said to be in linkage disequilibrium. Genetic polymorphisms can be of several types. The most common type is a single nucleotide polymorphism or SNP where a base is simply replaced by another. Insertions of additional sequences or deletions can also occur and these range in size from one to several thousand base pairs. Many genetic polymorphisms are silent with no effect on gene products but may still be useful genetic markers in studies on disease association. In general, functionally significant effects associated with genetic polymorphisms are most likely when they are associated with an amino acid substitution in the gene product, when a deletion or insertion results in a frameshift in the coding region, when a gene is completely deleted or when the polymorphism directly affects gene transcription, RNA splicing, mRNA stability or mRNA translation.

The availability of comprehensive sequence data from the Human Genome Project is likely to be followed later this year by the appearance of the 'SNP map' which will be a high-density map of 200 000–300 000 SNPs [2]. Since SNPs and other polymorphisms are useful markers for a variety of genetic studies including those on susceptibility to polygenic diseases and adverse drug reactions, it is suggested that the SNP map will facilitate more accurate drug prescribing and also the development of new drugs due to a better understanding of disease causation [2]. However, there is controversy as to the exact number of SNPs required for disease gene mapping and whether this will be feasible using random SNPs to map disease associations through linkage disequilibrium or whether only SNPs giving rise to functionally significant polymorphisms will provide meaningful information on complex diseases [3, 4]. There is already substantial information on SNPs and other polymorphisms available for many genes, especially those encoding enzymes important in drug metabolism and those affecting immunity and inflammation. A number of these are attractive candidates for disease susceptibility genes and this has already prompted a large number of case-control studies on associations between particular polymorphisms and diseases. Since improved understanding of the specific genes involved in disease causation is likely to lead to the development of new drug treatments, candidate gene association studies are very relevant to the area of clinical pharmacology. As information on candidate genes and candidate SNPs increases, it is important that study design and data analysis in disease-association studies utilizing them should receive careful attention to ensure that a large number of spurious associations are not reported and that genuine disease genes are identified.

Design and interpretation of disease association studies

Family studies and studies based on linkage-disequilibrium mapping

Almost all human diseases have some genetic component determining who will be affected though the extent of this component varies widely. The clearest example of a genetic disease is where there is a single defective gene which affects all individuals with the particular genotype. Examples of diseases that fall into this category include Huntington's disease, cystic fibrosis and certain types of inherited cancer. The normal approach to identifying the gene responsible for such disorders is to carry out linkage studies using large affected kindreds. Linkage studies differ in a number of ways from association studies as summarized in Table 1 though there is some overlap between the two types of study. In linkage studies, polymorphic genetic markers from different chromosomes (either microsatellite where a sequence of a few bp is repeated a variable number of times or SNP) are used to pinpoint the chromosomal location of the disease gene on the basis that markers close to the gene will cosegregate with the disease within the family.

Many common diseases ranging from cancer to Alzheimer's disease are polygenic, with more than one gene contributing to disease susceptibility in the population. In contrast to single gene disorders, in polygenic diseases possession of a particular variant of a gene

Table 1 Identification of disease genes.

Linkage studies
Used mainly up to the present in the study of single gene disorders
Use genetic markers situated throughout genome. These markers will generally not be functionally significant in terms of phenotypic effect
Use extended families or in certain types of study cases and unrelated controls
Linkage of genotype for a genetic marker to disease may be unique to the particular family
Association studies
Used to identify genes involved in polygenic disorders
Determine genotype for polymorphism (often functionally significant) in candidate gene of biological relevance to the disease
Use cases and appropriately matched controls who are normally unrelated
Association of a genotype or phenotype with disease is a statistical finding which is not necessarily a reflection of genetic linkage

associated with the disease does not mean that an individual will necessarily develop the disease, simply that they have an increased risk of disease development. Family based linkage studies can also be used in the identification of this type of gene but in this case, it is more usual to use large sets of different families [5]. One approach involves affected sib pair (ASP) analysis where genetic markers that are more common than expected in both affected sibs are identified [6]. In another approach known as the transmission disequilibrium test (TDT), markers preferentially transmitted to affected individuals from both their parents are identified [7]. The precise number of families required for TDTs depends on a factor known as a sibling recurrence risk ratio (λs). This ratio is calculated as the lifetime sibling risk divided by the population frequency. A second ratio known as λgs represents the risk ratio for individual genes and is calculated by partitioning λs between the number of genes hypothesized to contribute to the disease. If the λgs is less than 1.3 (i.e. the risk of siblings with a particular gene developing the disease is only slightly greater than the population frequency) and the allele frequency is less than 0.1, then successful use of TDT testing to pinpoint disease genes will require impractically large (>1000) numbers of families [8]. TDT testing can be used either for whole genome scanning where an area within a particular chromosome associated with the disease is identified on the basis of segregation pattern within families or for examining intrafamilial allelic association of genetic polymorphisms within candidate disease genes.

A major limitation of TDT testing is the need to collect DNA samples from parents of affected individuals. This is seldom a problem with diseases where onset is during childhood or early adulthood but the majority of common polygenic diseases such as Alzheimer's disease, cancer and noninsulin-dependent diabetes mellitus arise later in life increasing the likelihood that one or both parents will be deceased. With such diseases, there is also a risk that TDT testing will tend to identify genes that are important in rarer early onset forms of the disease since, in general, patients with two surviving parents will tend to be on average younger than those with both parents deceased. To overcome these limitations, variations on TDT testing using only siblings of the affected individual have recently been proposed [9]. The value of such tests is still under discussion but, as with the TDT, when λgs and the allele frequency of the candidate gene is low, achieving statistical significance without use of very large numbers of families is likely to be difficult.

With certain diseases such as alcoholism, alcoholic liver disease, AIDS and various forms of mental illness, family studies are hampered further by the associated stigma with the result that many affected individuals will not wish their parents or siblings to know of their illness. Studies in alcoholic liver disease are faced with two additional complications. First, unless other members of the family also abuse alcohol, it will not be possible to classify them as affected or unaffected and second, liver biopsy is ideally required to accurately determine the presence and severity of disease.

As an alternative to family studies, attempts are currently underway to detect novel genes associated with complex diseases by genome-wide SNP scanning to determine whether certain SNPs are more common in disease groups compared with controls. Though there is considerable enthusiasm for this approach, the number of SNPs needed for such a scan remains controversial and it has also been suggested that the unpredictable nature of genetic linkage may make finding real associations difficult [4].

Population candidate gene association studies

The alternative to family studies and to population approaches that rely on linkage disequilibrium is to perform population candidate gene association studies. As summarized in Table 1, there are a number of advantages to this approach including the fact that such studies may provide adequate power to detect relative risks as low as 1.5 which is usually not possible in family studies [3] and as most candidate gene studies are focusing directly on a single gene and frequently look directly at functionally significant polymorphisms, concerns about the extent of linkage disequilibrium and the adequacy of SNP markers to detect associations are not relevant [4]. In population candidate gene association studies, DNA samples from cases and population controls are genotyped for polymorphisms situated in or close to a gene which prior knowledge suggests might play a role in the pathogenesis of the disease of interest. Until recently, the polymorphisms studied were often genetic markers rather than polymorphisms giving a direct functional effect on the gene product. Increasing sequence information and more comprehensive studies on polymorphisms however, mean that it should usually now be possible to choose a functionally significant polymorphism in the gene of interest for study. There are a number of important issues to consider in the design of case-control studies of this type. These include: choice of candidate gene and polymorphism for study, recruitment methods, matching of controls and number of subjects to be studied.

Choice of candidate gene and polymorphism for study

The first step in designing a case-control study is to decide on the candidate gene or genes to be studied. Consideration should be given both to the relevance of the candidate gene to be studied to the pathogenesis of the particular disease and to the functional effects of a

particular polymorphism. As discussed recently by Risch [3], studying functionally significant polymorphisms rather than random polymorphisms in the gene of interest offers considerable advantages in terms of detecting disease-associated genes. It is also desirable that there should be a reasonable understanding of whether the polymorphism is linked to other polymorphisms either in the gene of interest or an adjacent gene. Functional effects of polymorphisms are complex and it is important that the overall effects of possession of a particular haplotype (combination of linked polymorphisms) be considered rather than the functional effects of a single polymorphism in isolation. Ideally, the functional significance of a polymorphism or a linked set of polymorphisms should be investigated both from the basis of the effect on *in vivo* phenotype and on function *in vitro*. As discussed below, it has been possible to study phenotype-genotype relationships in detail for a number of genes encoding enzymes of xenobiotic metabolism but in the case of other genes, this approach may not necessarily be feasible. The functional significance of polymorphisms resulting in amino acid substitutions in the gene product are most amenable to study *in vitro* providing suitable expression systems and assay methods for the protein of interest are available. There are more difficulties in studying the *in vitro* functional effects of polymorphisms in noncoding sequences. For example, as discussed below, studies on promoter region polymorphisms in the gene encoding the cytokine TNF-α using reporter gene constructs have yielded contradictory conclusions as to whether the polymorphism is functionally significant depending on the precise gene construct and cell line used in the experiments [10]. The availability of detailed information on functional significance before an association study is initiated should help to avoid the reporting of spurious associations and it seems likely that the availability of detailed information on human mutations [2] should lead to an improvement in this aspect of study design. In the past, many of the polymorphisms studied were those which gave rise to restriction fragment length polymorphisms (RFLPs) which were discovered by hybridization rather than by complete sequencing or mutation scanning studies. This may have led to functionally significant polymorphisms in candidate genes being ignored and almost certainly led to many negative studies looking for associations between diseases and polymorphisms that were neither functionally significant nor linked to a functionally significant polymorphism. A recent study described a detailed and systematic search for polymorphisms in genes relevant to hypertension [11]. However, the functional significance of these polymorphisms has not yet been determined and it is desirable that this be examined before undertaking disease association studies. Information of this type for all genes is likely to become increasingly available from SNP databases but again determining which polymorphisms are functional and concentrating on these may be more fruitful than random disease associations on all SNPs, though this remains controversial [3]. Approaches such as the analysis of phenotype in transgenic or knockout mice and studies on patterns of gene expression in diseased cells using microarray techniques may in future be helpful in deciding the most appropriate candidate genes for study.

Statistical aspects of study design

The frequency of the variant genotype to be studied will determine the number of cases and controls that need to be recruited to achieve sufficient statistical power to detect a difference in frequency between the two groups. For example, a typical study might aim for 80% statistical power to detect an association with a P value of less than 0.05 assuming an odds ratio of 2 for the variant genotype in disease development. If the variant genotype of interest occurs at a frequency of approx. 0.2 or less, the numbers of subjects required to see a statistically significant odds ratio may be unmanageably large particularly from the point of view of case and control recruitment unless the odds ratio is very high. Using twice or three times the number of controls compared with cases can be helpful where the number of cases is limited due to a disease being rare.

In practice, with the ever increasing volume of data on gene sequences and new polymorphisms, many recent studies involve genotyping cases and controls studied previously for other polymorphisms or a large number of different polymorphisms are studied concomitantly. In either case, consideration should be given to correcting for multiple testing especially where polymorphisms are being studied that have no functional significance or where their functional significance remains unclear. If this is not done, there is a risk that chance associations will be detected and this becomes more likely with increasing numbers of polymorphisms being studied. For example, a recent study on susceptibility to disease progression in patients with chronic hepatitis C infection examined eight different polymorphisms which, though present in genes relevant to disease progression, were largely of unproven functional significance, and reported statistically significant differences for two of the polymorphisms without any correction for multiple testing [12]. Had a Bonferroni correction been applied neither association would have been statistically significant. However, there is also a risk that use of multiple test corrections such as Bonferroni corrections may result in genuine associations being missed. A recent paper cautions against their indiscriminate use in all situations and suggests that assessing the likelihood of a chance association should be considered in the light of the biological plausibility of any observed

associations [13]. Bonferroni corrections will still be needed unless there is a preestablished hypothesis for the association being sought, which in most candidate gene association studies will require the candidate gene to be of clear biological relevance and the chosen polymorphism of proven functional significance.

Recruitment of cases and controls

Patient recruitment is a key factor in association studies. It is important that this should be organized so that new incident cases are recruited soon after diagnosis rather than relying on recruiting patients already diagnosed. If preexisting cases are enrolled, there is a risk that the genetic factor may be studied from the point of view of disease progression and severity rather than risk of disease development. The recruitment rate of eligible patients should also be as high as possible to ensure that there is no risk of obtaining a significant association relating to willingness to participate in research rather than with the disease itself. It is also important that appropriate and uniform diagnostic criteria for the disease being studied are used, especially if subgroup analysis involving the stage or severity of the disease is to be performed. Since lifestyle and occupational factors play an important part in determining susceptibility to many diseases, it is important that appropriate questionnaires to cover these factors are designed and that arrangements are in place to ensure they are completed by all subjects. There are also many pitfalls in control recruitment. It is often convenient to use hospital staff or patients suffering from another disease as controls. This is not good practise. Ideally, cases and controls should be matched by ethnic origin, age and gender and this can normally only be done by seeking community-based controls. In addition, with some types of diseases, it is important that additional matching be performed. For example, in a study on susceptibility to alcoholic liver disease, alcoholics with a similar drinking history to the controls are a particularly appropriate control group. In the case of lung cancer, where the risk of tobacco smoking is likely to be higher than any genetic factor, it is helpful to use smokers as controls though statistical corrections for the effect of smoking can also be made.

A frequently cited concern with population disease association studies is that the disease population will not necessarily be representative of the population at large due to a 'founder effect' or 'population stratification' where the disease gene originated in a single kindred who settled in the area. A new approach to this problem involves the use of method known as 'genomic control' where the frequency of polymorphisms in other genes unlikely to be associated with the disease are compared between subjects in addition to the polymorphism in the candidate gene of interest and appropriate adjustment is then made for population stratification [14]. However, a recent study on population stratification in US Caucasians of European origin has concluded that odds ratio calculations are biased by less than 10% except under extreme conditions for this population group suggesting that the danger of population stratification leading to unreliable data may have been overemphasized by some workers [15].

In addition to the possibility of spurious associations due to population stratification, associations may arise because of other biases in relation to case and/or control recruitment. An approach to dealing with concerns about case and control recruitment is to replicate the study at a completely different centre. Failure to replicate findings in a different ethnic group may not mean that the original study is flawed. Differences in allele frequency between ethnic groups means that if a risk factor is seen at a lower frequency in another ethnic group, there may no longer be sufficient statistical power to detect a population association though individuals with the genotype of interest could be still at risk. Similarly if gene–environment interactions are important in determining the magnitude of a genetic risk factor, another ethnic group may not normally be exposed to the same environmental factor as in the original study and the gene may therefore be irrelevant in the second ethnic group. It is particularly helpful in confirming findings from a case–control disease association study if the findings can be replicated in families using TDT or other appropriate tests but, as discussed previously, this is not always feasible.

Examples of case-control studies relevant to clinical pharmacology

Studies on genes affecting drug response

There have now been a number of published studies where the relationship between genotype for a particular drug target or metabolizing enzyme and drug response or toxicity is examined with view to using genotype data to determine the most appropriate drug dosage or whether prescribing the drug is likely to be effective. In general, most of these studies have involved relatively small group sizes and, in some cases, there are problems with inadequate control groups. Many of these studies are not true case-control studies in that they compare outcome between two groups of affected individuals rather than between cases and controls. However, many of the requirements for a good case-control study such as those relating to recruitment and statistical analysis will also be valid in this type of study.

It has been suggested that Alzheimer's disease patients positive for at least one ApoE4 allele show a poorer response to tacrine compared with those with no ApoE4

alleles [16]. However, a number of follow-up studies mainly involving larger group sizes have failed to confirm these observations with one study reporting that the difference between genotypes could only be confirmed in women and another reporting that ApoE4 positivity was associated with a better response to the drug [17, 18]. The inconsistencies may be due to the actual differences in response between groups being relatively small but could also reflect the fact that there appear to be ethnic differences in the consequences of the ApoE4 allele [19].

Genotyping for polymorphisms in other genes concerned with drug response or metabolism have also been suggested to affect response or toxicity. A codon 9 polymorphism in the dopamine receptor DRD3 gene appears to predict risk of tardive dyskinesia development in patients receiving neuroleptic drugs [20]. However, a follow-up study failed to obtain a statistically significant increased risk though the overall trend was similar to that observed in the original study [21].

Asthma patients with no wild-type alleles for an upstream repeat polymorphism in the 5-lipoxygenase gene appear unresponsive to treatment with a 5-lipoxygenase inhibitor though the fact that the numbers of patients with the unresponsive genotype was considerably lower than the number in the wild-type group is a limitation of the study [22]. In other studies of drug responses in asthma patients, it has been observed that genotype for certain β2 adrenergic receptor polymorphisms predicts bronchodilatory response to β-adrenoceptor agonists together with tachyphylaxis to repeated use, but the overall results from all the studies are not consistent [23–25]. It has now been demonstrated that haplotype for a total of 13 SNPs in the upstream and coding sequences of the β_2-adrenergic receptor appears to be a better predictor of agonist response with five different haplotypes relatively common among Caucasians [26]. In the future, genotyping methods which determine haplotype for β_2-adrenergic receptor rather than genotypes for single SNPs may therefore be of value but determining haplotype is currently a complex process.

There have been a number of studies concerned with identifying genotypes which can predict clozapine response. A recent study focused on 19 different receptor and transporter polymorphisms and detected significant associations for six different genotypes. It was suggested that genotyping for two SNPs in the 5-HT2A receptor would predict clozapine response in most patients [27]. However, these findings were not confirmed in another study [28] and no corrections for multiple testing were applied in spite of the fact that the functional significance of many of the polymorphisms studied is still unclear.

In the case of the cytochromes P450 (CYP) CYP2D6, CYP2C9 and CYP2C19, it is now clear that genotype will have important effects on drug response and toxicity with some substrates. Though the existence of the CYP2D6 polymorphism is well established, there have been relatively few association studies in relation to adverse drug reactions performed. However, two polymorphisms affecting activity of CYP2C9, the major *S*-warfarin metabolizing enzyme, have been well characterized and have recently been examined in relation to warfarin dose requirement and risk of bleeding. By comparing CYP2C9 genotypes in a group of patients with an unusually low warfarin dose requirement with both a group of random patients from the same clinic and a local control group, it was found that possession of one or two variant CYP2C9 alleles was significantly associated with a low (1.5 mg day^{-1} or less) dose requirement and an increased risk of bleeding [29]. The findings in this relatively small study have subsequently been confirmed in a larger study involving 561 patients [30].

Case-control studies on genes encoding enzymes of xenobiotic metabolism

Because of their general role in xenobiotic biotransformation, genes encoding enzymes such as the cytochromes P450, glutathione *S*-transferases (GST) and *N*-acetyltransferases (NAT) are good candidates for susceptibility factors in many diseases. There have been a large number of case-control studies on the relationship of polymorphisms in these genes to susceptibility to cancer and some other diseases such as Parkinson's disease [31]. The majority of polymorphisms in these genes studied in disease-association studies are of clear functional significance; studies on functional significance have been facilitated by previous studies on phenotype and by the availability of well characterized expression systems and sensitive enzyme assays. Despite these advantages, there have been some problems with the interpretation of studies. The most positive and consistent associations reported are in relation to an absence of GSTM1 activity and a high NAT2 activity being risk factors for lung cancer and colorectal cancer, respectively [32, 33]. However, both associations are not always obvious from the overall case-control data. Subgroup analysis in relation to the risk factors of smoking history for lung cancer and meat consumption and cooking method for colorectal cancer gives the most significant data, illustrating the importance of gene–environment interactions [34–36].

There have been more problems with studies involving the cytochromes P450 CYP2D6 and CYP1A1 in relation to lung cancer susceptibility. In the case of CYP2D6, an initial study suggested that a high level of enzyme activity was a risk factor for the development of lung cancer and this was confirmed by a later study [37, 38]. However, recent larger and better designed studies have shown no apparent association and meta analysis also suggests the

absence of a significant effect [39–42]. In view of what is now known about the substrate specificity of CYP2D6 [43] and the fact that it is apparently not expressed in the lung [44], the recent findings on CYP2D6 and lung cancer are not surprising and CYP2D6 now seems a poor candidate gene for this disease.

In the case of CYP1A1, an apparently consistent association between homozygosity for two linked polymorphisms (designated the *CYP1A1*2* allele) and susceptibility to lung cancer in Japanese has been reported [45, 46]. However, the functional significance of the two linked polymorphisms is unclear. One results in an amino acid substitution but effects on enzyme activity appear relatively small [47–49] and the second is some distance downstream of the coding sequence. It remains possible that this is functionally significant. Alternatively, there may be another linked polymorphism in Orientals that directly affects enzyme activity or gene expression though there is no evidence for this from existing sequence data. The *CYP1A1*2* allele is rarer in Caucasians than Orientals and, in line with this, population association studies in Causasians on this allele have shown either no association or only borderline significance. Since CYP1A1 is expressed in lung tissue exposed to inducers such as benzo[*a*]pyrene and is the main activating enzymes for several polycyclic aromatic hydrocarbons, this gene remains a good candidate for modulating lung cancer risk [50, 51]. The reported association of *CYP1A1*2* with the disease needs to be confirmed by obtaining reliable data on the functional significance of the 3′-end polymorphism or by identification of a functionally significant linked polymorphism either in *CYP1A1* itself or a neighbouring gene.

Another cytochrome P450, CYP2E1, has been well studied in association studies on alcoholic liver disease because of its role in ethanol oxidation and the presence of upstream polymorphisms. There is some evidence that a rare variant allele detectable with the restriction enzyme *Rsa*I may be associated with an increased risk for development of alcoholic liver disease and an earlier age of onset of this disease [52, 53]. Some problems with the interpretation of these observations remain. In particular, the functional significance of the polymorphism remains controversial and the low frequency of the rare allele in Caucasians (0.015) means that it is not an important population risk factor for the disease. Though CYP2E1 is inducible by alcohol, it is not the major oxidizing enzyme in the liver even when induced and oxidizes the toxic initial product acetaldehyde as well as ethanol [54]. However, since ethanol oxidation by CYP2E1 appears to be also associated with free radical formation which may contribute to the onset of alcoholic liver disease [55], variant alleles associated with high activity may be minor population risk factors for alcoholic liver disease. Attempts to identify more common alleles in this gene which give functionally significant effects have been largely unsuccessful [56, 57].

CYP2A6 is the main cytochrome P450 isoform that contributes to nicotine metabolism, converting nicotine to cotinine. The description of two apparently functionally significant polymorphisms in this gene several years ago prompted a number of studies to determine whether decreased CYP2A6 activity might be associated with altered smoking behaviour due to impaired ability to metabolize nicotine. One study suggested that possession of one or two variant alleles decreased the likelihood of smoking and also the number of cigarettes smoked though this could not be confirmed in two other studies [58–60]. Unfortunately, recent studies on the CYP2A6 gene have demonstrated that the assays used in these genotyping studies could lead to artefactual results and indeed one of the two variant alleles, *CYP2A6*3*, does not appear to exist *in vivo* [61]. More reliable genotyping methods for CYP2A6 are now available and have recently been used for a further study which largely confirmed the earlier findings in one study of a link between CYP2A6 genotype and smoking behaviour [62]. However, the variant CYP2A6 alleles are rarer in Caucasians than initially described and are therefore unlikely to be major contributors to determining who is likely to become addicted to nicotine. These difficulties in establishing reliable methods for CYP2A6 genotyping are mainly due to the existence of a number of highly homologous CYP2A genes but illustrate the importance of validating genotyping assays before embarking on molecular epidemiological studies.

Case-control studies on immunogenetic polymorphisms

Many common polygenic diseases ranging from insulin-dependent diabetes mellitus to multiple sclerosis have an autoimmune component. Interindividual variation in genes encoding proteins with an involvement in the immune and inflammatory responses is therefore a potentially important susceptibility factor. There is now increasing information available on genetic polymorphisms in such genes, particularly those in the MHC class II complex and those encoding cytokines such as tumour necrosis factor-α (TNF-α), interleukin-10 (IL-10) and interleukin-4 (IL-4). Particular MHC class I and class II haplotypes have been shown to be associated with a number of different diseases. In many cases such as insulin-dependent diabetes mellitus and rheumatoid arthritis, these are strong and well established associations, confirmed by a large number of studies though some other suggested associations remain controversial [5]. The complexity and the large extent of polymorphism in the MHC locus means that all associations with particular haplotypes need

to be corrected for multiple testing. The role of the MHC complex in antigen presentation to T cells is very well established but the functional significance of the various haplotypes is still not fully clear making interpretation of associations difficult.

Disease association studies on polymorphisms in cytokine genes and cytokine receptor genes have also been reported. Thus far, the cytokine gene that has received the most attention is TNF-α, possibly because it is located within the MHC gene cluster, raising the possibility that some reported MHC associations might be due to linkage to particular TNF-α alleles. There are a number of different polymorphisms in the promoter region of TNF-α and also at least one situated in the coding region but the most frequent and also the best studied from the point of view of disease association are those at positions −238 and −308. Though associations of these polymorphisms with suscepibility to a variety of diseases have been reported, consistent results have not been obtained with some positive and some negative associations being reported. The failure to see consistent associations might in part be due to lack of functional significance for these polymorphisms. As discussed by Allen [10], there have been a number of studies on the functional significance of both polymorphisms with some reporting increased transcriptional activity for the variant alleles and some no difference from the wild-type. In the absence of reliable and consistent functional data, it is possible that many of the reported associations between TNF-α alleles and disease susceptibility are either chance associations or reflect linkage with MHC genes.

IL-10 has an important role as a regulator of T cell responses and also has general anti-inflammatory effects. Several upsteam polymorphisms, both SNPs and microsatellites, occur in this gene. In the case of a polymorphism at position −627, there is evidence from both reporter gene studies and *in vivo* measurements and reporter gene studies that the variant allele is associated with decreased IL-10 transcription [63, 64]. The presence of the variant allele appears to increase the risk for fibrotic alcoholic liver disease in alcoholics, and is associated with more severe forms of juvenile rheumatoid arthritis [65, 66]. However, hepatitis C patients homozygous for the IL-10 variant show a better response to interferon-α treatment and a lower rate of disease progression, possibly due to the higher IL-10 levels associated with the wild-type allele interfering with an effective antiviral response [12, 67, 68]. There appear to be problems in selecting the most appropriate IL-10 polymorphism for study. Though there is evidence that it is the C-627 A polymorphism that directly affects IL-10 transcription [63], several studies have focused on either the A-1117G polymorphism or the microsatellite polymorphisms which, although linked to −627, are not in complete linkage disequilibrium with this polymorphism [69, 70]. As 50% of individuals positive for G-1117 will have the wild-type C at position −627 and secrete normal levels of IL-10, the failure to detect strong associations in some studies could relate to failure to examine the most appropriate polymorphism.

In addition to cytokine genes, genotype for polymorphisms in cytokine receptors may also be useful predictors for susceptibility to autoimmune disease. There is increasing evidence supporting a role for IL-4 receptor (IL4R) genotype in predicting susceptibility to atopy and asthma and such an association is biologically sensible in view of the role of IL-4 in regulation of Th2 immune responses and IgE secretion. A number of different polymorphisms in IL4R have been described and associations with atopy and asthma reported [71, 72]. However, the results have been inconsistent with different studies reporting different polymorphisms as being significant for disease susceptibility. A recent study appears to explain at least some of these consistencies. It is now clear that there are six different polymorphisms in the coding region of IL4R which result in amino acid substitutions [73]. There is a complex pattern of linkage between the various polymorphisms resulting in a range of haplotypes. It appears that susceptibility to atopy may relate particularly to the presence of a polymorphism giving rise to the amino acid substitution Cys406Arg but overall haplotype for all six polymorphisms may also contribute [73]. As with the β-adrenergic receptor gene discussed previously [26], genotyping for six separate polymorphisms and assigning haplotypes is technically complex. Though there is currently some evidence that two of the amino acid substitutions are functionally significant, there is a need for additional studies to determine precisely which IL4R haplotypes result in altered biological activity to facilitate additional case-control studies on the relationship of IL4R genotype to disease susceptibility. A similar situation is likely to arise with other candidate genes as more SNPs are discovered.

Concluding remarks

There are clear potential pitfalls in both the design and interpretation of candidate gene association studies but the availability of more comprehensive information about genetic polymorphisms and the increasing availability of better systems for studying functional effects makes it likely that such studies will continue and will in the future provide useful information on specific gene defects in disease. Providing findings can be replicated in more than one study population, associations (or indeed lack of assocation) should provide important information about disease mechanisms and the ability to predict individuals who are at risk. Such associations may also lead to the design of new treatments that target gene defects more

directly potentially leading to improvements in public health. However, there are considerable ethical problems associated with the identification of individuals at increased risk of a disease, especially where there is no effective preventive treatment available, and these issues need careful discussion so that genetic information can be used for individual benefit.

References

1 Vogel F, Motulsky AG. *Human genetics. Problems and Approaches*, 3rd edn. Berlin. Springer-Verlag, 1996.

2 Roses AD. Pharmacogenetics and the practise of medicine. *Nature* 2000; **405**: 857–865.

3 Risch NJ. Searching for genetic determinants in the new millenium. *Nature* 2000; **405**: 847–856.

4 Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nature Genet* 2000; **26**: 151–157.

5 Strachan T, Read AP. Genetic mapping of complex characters. In *Human molecular genetics*, 2nd edn. New York. Wiley-Liss., 1999: 283–294.

6 Kruglyak L, Lander ES. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet* 1995; **57**: 439–454.

7 Speilman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium. *Am J Hum Genet* 1996; **59**: 983–989.

8 Scott WK, Pericak-Vance MA, Haines JL. Genetic analysis of complex diseases. *Science* 1997; **275**: 1327.

9 Speilman RS, Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* 1998; **62**: 450–458.

10 Allen RD. Polymorphisms of the human TNF-α promoter-random variation or functional diversity. *Mol Immunol* 1999; **36**: 1017–1027.

11 Halushka MK, Fan J-B, Bentley K, *et al*. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet* 1999; **22**: 239–247.

12 Powell EE, Edwards-Smith CJ, Hay JL, *et al*. Host genetic factors influence disease progression in chronic hepatitis C. *Hepatology* 2000; **31**: 828–833.

13 Perneger TV. What's wrong with Bonferroni adjustments? *Br Med J* 1998; **316**: 1235–1237.

14 Bacanu S-A, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; **66**: 1933–1944.

15 Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiological studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000; **92**: 1151–1158.

16 Poirier J, Delile M-C, Quirion R, *et al*. Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outocme in Alzheimer disease. *Proc Natl Acad Sci USA* 1995; **92**: 12260–12264.

17 Farlow MR, Lahiri DK, Poirier J, Davignon J, Schneider L, Hui SL. Treatment outcome of tacrine therapy depends upon apolipoprotein genotype and gender of the subjects with Alzheimer's disease. *Neurology* 1998; **50**: 669–677.

18 Rigaud AS, Traykov L, Caputo L, *et al*. The apolipoprotein E epsilon 4 allele and the response to tacrine therapy in Alzheimer's disease. *Eur J Neurol* 2000; **7**: 255–258.

19 Tang MX, Maestre G, Tsai WY, *et al*. Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. *Am J Hum Genet* 1996; **58**: 574–584.

20 Steen VM, Lovlie R, MacEwan T, McCreadie RG. Dopamine D3-receptor gene variant and susceptibility to tardive dyskinesia in schizophrenic patients. *Mol Psychiatry* 1996; **2**: 139–146.

21 Lovlie R, Daly AK, Blennerhassett R, Ferrier N, Steen VM. Homozygosity for the Ser9Gly variant of the dopamine D3 receptor and risk for tardive dyskinesia in schizophrenic patients: a follow-up study. *Int J Neuropsychopharmacol* 2000; **3**: 61–65.

22 Drazen JM, Yandava CN, Dube L, *et al*. Pharmacogenetic association between *ALOX5* promoter genotype and the response to anti-asthma treatment. *Nature Genet* 1999; **22**: 168–170.

23 Martinez FD, Graves PE, Baldini M, Solomon S, Erickson R. Association between genetic polymorphisms of the beta (2)-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J Clin Invest* 1997; **100**: 3184–3188.

24 Tan S, Hall IP, Dewar J, Dow E, Lipworth B. Association between beta (2)-adrenoceptor polymorphism and susceptibility to bronchodilator desensitisation in moderately severe stable asthmatics. *Lancet* 1997; **350**: 995–999.

25 Israel E. The effect of polymorphisms of the beta (2)-adrenergic receptor on the response to regular use of albuterol in asthma. *Am J Respir Crit Care Med* 2000; **162**: 75–80.

26 Drysdale CM, McGraw DW, Stack CB, *et al*. Complex promoter and coding region beta (2)-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci USA* 2000; **97**: 10483–10488.

27 Arranz MJ, Munro J, Birkett J, *et al*. Pharmacogenetic prediction of clozapine response. *Lancet* 2000; **355**: 1615–1616.

28 Schumacher J, Schulze TG, Wienker TF, Rietschel M, Nothen MM. Pharmacogenetics of clozapine response. *Lancet* 2000; **356**: 506–507.

29 Aithal GP, Day CP, Kesteven PJL, Daly AK. Relationship of polymorphisms in the cytochrome P450 CYP2C9 to warfarin dose requirements and risk of bleeding complications. *Lancet* 1999; **353**: 717–719.

30 Taube J, Halsall D, Baglin T. Influence of cytochrome P-450CYP2C9 polymorphisms on warfarin sensitivity and risk of over-anticoagulation in patients on long-term treatment. *Blood* 2000; **96**: 1816–1819.

31 Smith G, Stanley LA, Sim E, Strange RC, Wolf CR. Metabolic polymorphisms and cancer susceptibility. *Cancer Surveys* 1995; **25**: 27–65.

32 McWilliams JE, Sanderson BJS, Harris EL, Richert-Boe KE, Henner WD. Glutathione *S*-transferase M1 (GSTM1) deficiency and lung cancer risk. *Cancer Epidemiol Biomarkers Prev* 1995; **4**: 589–594.

33 d'Errico A, Malats N, Vineis P, Boffetta P. Review of studies of selected polymorphisms and cancer. In *Metabolic Polymorphisms and Susceptibility to Cancer,* eds. Vineis P, Malats N, Lang M, *et al*. Lyon, France, IARC Scientific Publications, 1999: 323–393.

34 London SJ, Daly AK, Cooper J, Navidi WC, Carpenter CL, Idle JR. Polymorphism of glutathione *S*-transferase M1 (GSTM1) and risk of lung cancer among African-Americans and Caucasians in Los Angeles county. *J Natl Cancer Inst* 1995; **87**: 1246–1253.

35 Roberts-Thompson IC, Ryan P, Khoo KK, Hart WJ, McMichael AJ, Butler RN. Diet, acetylator phenotype and risk of colorectal neoplasia. *Lancet* 1996; **347**: 1372–1375.

36 Welfare MR, Cooper J, Bassendine MF, Daly AK. Relationship between acetylator status, smoking, diet and colorectal cancer risk in the north-east of England. *Carcinogenesis* 1997; **1997**: 1351–1354.

37 Ayesh R, Idle JR, Ritchie JC, Crothers MJ, Hetzel MR. Metabolic oxidation phenotypes as markers for susceptibility to lung cancer. *Nature* 1984; **311**: 169–170.

38 Caporaso NE, Tucker MA, Hoover RN, *et al*. Lung cancer and the debrisoquine metabolic phenotype. *J Natl Cancer Inst* 1990; **82**: 1264–1272.

39 Wolf CR, Smith CAD, Gough AC, *et al*. Relationship between the debrisoquine hydroxylase polymorphism and cancer susceptibility. *Carcinogenesis* 1992; **13**: 1035–1038.

40 Stucker I, Cosme J, Laurent P, *et al*. CYP2D6 genotype and lung cancer risk according to histologic type and tobacco exposure. *Carcinogenesis* 1995; **16**: 2759–2764.

41 London SJ, Daly AK, Leathart JBS, Navidi WC, Carpenter CC, Idle JR. Genetic polymorphism of *CYP2D6* and lung cancer risk in African-Americans and Caucasians in Los Angeles county. *Carcinogenesis* 1997; **18**: 1203–1214.

42 Rostami-Hodjegan A, Lennard MS, Woods HE, Tucker GT. Meta-analysis of studies of the CYP2D6 polymorphism in relation to lung cancer and Parkinson's disease. *Pharmacogenetics* 1998; **8**: 227–238.

43 Patten CJ, Smith TJ, Murphy SE, Wang MH, Lee J, Tynes RE, Koch P, Yang CS. CYP2D6 and NNK kinetic analysis of the activation of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone by heterologously expressed human P450 enzymes and the effect of P450-specific chemical inhibitors on this activation in human liver microsomes. *Arch Biochem Biophys* 1996; **333**: 127–138.

44 Kivisto KT, Griese E-U, Stuven T, *et al*. Analysis of CYP2D6 expression in human lung: implications for the association between CYP2D6 activity and susceptibility to lung cancer. *Pharmacogenetics* 1997; **7**: 295–302.

45 Kawajiri K, Nakachi K, Imai K, Yoshii A, Shinoda N, Watanabe J. Identification of genetically high risk individuals to lung cancer by DNA polymorphisms of the cytochrome P450IA1 gene. *FEBS Lett* 1990; **263**: 131–133.

46 Hayashi S, Watanabe J, Nakachi K, Kawajiri K. Genetic linkage of lung cancer-associated *Msp*I polymorphisms with amino acid replacement in the heme binding region of the human cytochrome P450IA1 gene. *J Biochem* 1991; **110**: 407–411.

47 Zhang Z, Fasco MJ, Huang L, Guengerich F, Kaminsky L. Characterization of purified human recombinant cytochrome P450-Ile462 and -Val462: Assessment of a role for the rare allele in carcinogenesis. *Cancer Res* 1996; **56**: 3929–3933.

48 Persson I, Johansson I, Ingelman-Sundberg M. In vitro kinetics of two human CYP1A1 variant enzymes suggested to be associated with interindividual differences in cancer susceptibility. *Biochem Biophys Res Commun* 1997; **231**: 227–230.

49 Schwarz D, Kisselev P, Cascorbi I, Schunck W-H, Roots I. Differential metabolism of benzo[*a*]pyrene and benzo[*a*]pyrene-7,8-dihydrodiol by human CYP1A1 variants. *Carcinogenesis* 2001; **22**: 453–459.

50 McLemore TL, Adelberg S, Liu MC, *et al*. Expression of CYP1A1 gene in patients with lung cancer: Evidence for cigarette-smoke induced gene expression in normal lung tissue and for altered gene regulation in primary pulmonary carcinomas. *Cancer Res* 1990; **82**: 1333–1339.

51 Bauer E, Guo Z, Ueng Y, Bell C, Zeldin D, Guengerich FP. Oxidation of benzo[a]pyrene by recombinant human cytochrome P450 enzymes. *Chem Res Toxicol* 1995; **8**: 136–142.

52 Pirmohamed M, Kitteringham NR, Quest LJ, *et al*. Genetic polymorphism of cytochrome P4502E1 and risk of alcoholic liver disease in Caucasians. *Pharmacogenetics* 1995; **5**: 351–357.

53 Grove J, Brown ASM, Daly AK, Bassendine MF, James OFW, Day CP. The *Rsa*I polymorphism of CYP2E1 and susceptibility to alcoholic liver disease in Caucasians: effect of age of presentation and dependence on alcohol dehydrogenase genotype. *Pharmacogenetics* 1998; **8**: 335–342.

54 Bell-Parikh LC, Guengerich FP. Kinetics of cytochrome P450 2E1-catalyzed oxidation of ethanol to acetic acid via acetaldehyde. *J Biol Chem* 1999; **274**: 23833–23840.

55 Albano E, Tomasi A, Persson JO, Terelius Y, Goriagatti L, Ingelman-Sundberg M, Dianzani MU. Role of ethanol-inducible cytochrome P450 (P450IIE1) in catalyzing the free radical activation of aliphatic alcohols. *Biochem Pharmacol* 1991; **41**: 1895–1902.

56 Hu Y, Oscarson M, Johansson I, *et al*. Genetic polymorphism of human *CYP2E1*: characterization of two variant alleles. *Mol Pharmacol* 1997; **51**: 370–376.

57 Fairbrother KS, Grove J, de Waziers I, *et al*. Detection and characterization of novel polymorphisms in the CYP2E1 gene. *Pharmacogenetics* 1998; **8**: 543–552.

58 Pianezza ML, Sellers EM, Tyndale RF. Nicotine metabolism defect reduces smoking. *Nature* 1998; **393**: 750.

59 London SJ, Idle JR, Daly AK, Coetzee GA. Genetic variation of CYP2A6, smoking, and risk of cancer. *Lancet* 1999; **353**: 898–899.

60 Sabol SZ, Hamer DH. An improved assay shows no association between the CYP2A6 gene and cigarette smoking behavior. *Behav Genet* 1999; **29**: 257–261.

61 Oscarson M, McLellan RA, Gullstén H, *et al*. Identification and characterisation of novel polymorphisms in the *CYP2A* locus: implications for nicotine metabolism. *FEBS Lett* 1999; **460**: 321–327.

62 Rao YS, Hoffmann E, Zia M, *et al*. Duplications and defects in the CYP2A6 gene: Identification, genotyping, and in vivo effects on smoking. *Mol Pharmacol* 2000; **58**: 747–755.

63 Rosenwasser LJ, Borish L. Genetics of atopy and asthma: The rationale behind promoter-basedcandidate gene studies (IL-4 and IL-10). *Am J Respir Crit Care Med* 1997; **156**: S152–S155.

64 Turner DM, Williams DM, Sankaran D, Lazarus M, Sinnott PJ, Hutchinson IV. An investigation of polymorphism in the interleukin-10 gene promoter. *Eur J Immunogenet* 1997; **24**: 1–8.

65 Grove J, Daly AK, Bassendine MF, Gilvarry E, Day CP. Interleukin 10 promoter region polymorphisms and susceptibility to advanced alcoholic liver disease. *Gut* 2000; **46**: 540–545.

66 Crawley E, Kay R, Sillibourne J, Patel P, Hutchinson I, Woo P. Polymorphic haplotypes of the interleukin-10 5′ flanking region determine variable interleukin-10 transcription and are associated with particular phenotypes of juvenile rheumatoid arthritis. *Arthritis Rheum* 1999; **42**: 1101–1108.

67 Edwards-Smith CJ, Jonsson JR, Purdie DM, Bansal A, Shorthouse C, Powell EE. Interleukin-10 promoter polymorphism predicts initial response of chronic hepatitis C to interferon alfa. *Hepatology* 1999; **30**: 526–530.

68 Yee LJ, Tang J, Gibson AW, Kimberly R, van Leeuwen DJ, Kaslow RA. Interleukin 10 polymorphisms as predictors of sustained response in antiviral therapy for chronic hepatitis C infection. *Hepatology* 2001; **33**: 708–712.

69 Maurer M, Kruse N, Giess R, Toyka KV, Rieckmann P. Genetic variation at position-1082 of the interleukin 10 (IL10) promotor and the outcome of multiple sclerosis. *J Neuroimmunol* 2000; **104**: 98–100.

70 Eskdale J, Wordsworth P, Bowman S, Field M, Gallagher G. Association between polymorphisms at the human IL-10 locus and systemic lupus erythematosus. *Tissue Antigens* 1997; **49**: 635–639.

71 Hershey GKK, Friedrich MF, Esswein LA, Thomas ML, Chatila TA. The association of atopy with a gain-of-function mutation in the alpha subunit of the interleukin-4 receptor. *New Engl J Med* 1997; **337**: 1720–1725.

72 Mitsuyasu H, Izuhara K, Mao XQ, *et al*. Ile50Val variant of IL4R alpha upregulates IgE synthesis and associates with atopic asthma. *Nature Genet* 1998; **19**: 119–120.

73 Ober C, Leavitt SA, Tsalenko A, *et al*. Variation in the interleukin 4-receptor α gene confers susceptibility to asthma and atopy in ethnically diverse populations. *Am J Hum Genet* 2000; **66**: 517–526.