Candidate gene studies

# Candidate gene studies in respiratory disease: avoiding the pitfalls

I Hall

Association studies of candidate gene polymorphisms in respiratory disease are easy to perform, but there are many pitfalls which have led to considerable criticism of these studies.

With the completion of a working draft of the human genome sequence, we are standing at the threshold of the most exciting period in biomedical research. It is now clear that the human genome contains around 32 000 genes,[1] and information is rapidly accumulating on the degree of interindividual variability in the sequences of these genes through collaborations such as the single nucleotide polymorphism (SNP) consortium.[2] Differences between individuals in their susceptibility to develop common diseases, the severity of their disease, and individual responses to treatment are all potentially predictable on the basis of the interaction of environmental factors with an individual's particular set of genotypes. Over the last few years this has led to a proliferation of association studies examining candidate gene polymorphisms for their potential role in respiratory disease. These studies are easy to perform—needing only a stored DNA sample and access to a relevant clinical data set—but there are many potential pitfalls to performing candidate gene association studies which have resulted in substantial criticism of this kind of study. This editorial provides guidance on some aspects of the design of such studies.

Some of the major issues which need to be considered in the design of candidate gene association studies are listed in table 1 and are briefly discussed below. It is estimated that SNPs occur at a rate of approximately 1 in 300 base pairs in non-coding DNA and 1 in 600 base pairs in coding DNA. Given that most genes are at least 1 kb in length (and many are much greater), it is immediately apparent that, even considering only the coding regions within the genome, there are very large numbers of polymorphisms available for study. Although the majority of polymorphic variation is accounted for by SNPs of which two million are currently described in the public databases, other polymorphisms do occur less frequently including deletions and insertions, both of which may introduce frame shifts. Because recombination is rare between polymorphisms which are close together in genetic terms—for example, two SNPs within the same gene—association studies are further complicated by the phenomenon of linkage disequilibrium. Linkage disequilibrium results in combinations of polymorphisms at a given locus (for example, within a gene and its regulatory region) occurring more frequently than would be predicted by chance alone. This causes problems in association studies because, if association is seen with a SNP at a given position, the basis for that association (assuming that it is real) may not be the SNP which has been studied but may be another polymorphism nearby which is in linkage disequilibrium with the SNP which has been genotyped. One approach to dealing with this is to look at associations with haplotypes—that is, combinations of alleles at different polymorphisms within a locus.

**Table 1** Problems with genotype patient studies

| Area of concern | Solution |
|---|---|
| Is there a clear hypothesis? | Use functional data if available |
| Has population stratification been allowed for? | Use individuals drawn from the same population/same racial group |
| Is the study adequately powered? | Calculate in advance the numbers needed in each group depending upon the allelic frequency for the polymorphism(s) under consideration, the number of phenotypes to be studied, the number of polymorphisms to be studied and the likely magnitude of effect of any polymorphism |
| Is linkage disequilibrium a problem? | Study several SNPs within a given gene and consider whether a haplotype analysis may help |
| Is multiple testing on this population an issue? | State the number of studies performed in the population previously and adjust the statistical analysis accordingly |

The major criticism which has been levelled at most candidate gene studies is that they are underpowered, particularly where several end points are being examined.[3] Few functional data are available on the SNPs which have been described to date in the human genome, so the majority of association studies merely use SNPs as markers. If multiple SNPs are being examined for association within a population with multiple phenotypes (or end points), then correction must be made for multiple testing. Surprisingly, many studies submitted to *Thorax* and other journals fail to allow for this. If functional data are available for a specific polymorphism enabling a hypothesis to be generated before the study is performed, then this may help in reducing the number of subjects required for the study by reducing the number of comparisons to be made in the analysis. Being cynical, too many of the studies which have been published to date have generated hypotheses to fit the data after the study is complete in order to justify the relatively small populations. Many SNPs do not alter the amino acid sequence of the relevant gene product and hence would be unlikely to have marked functional effects, but some polymorphisms are likely to give rise to functional effects. One example would be the CCR5Δ32 mutation: CCR5 is a chemokine receptor which is also important as a co-receptor for HIV entry into cells.[4] Individuals homozygous for the Δ32 deletion (which introduces a frame shift and results in the production of a truncated protein which is not expressed at the cell surface) are essentially human "knock outs" for CCR5 and hence would be hypothesised to be relatively resistant to HIV infection following exposure.[5] This has been shown in several clinical studies. Knowledge of the functional consequences of this mutation could be used to define the specific phenotype which would be predicted to be associated with CCR5Δ32 in an association study.

Another problem in assessing functionality is predicting what the functional consequences in homozygous and heterozygous individuals might be. Because we all have two copies of all autosomal genes, an individual may be homozygous or heterozygous for any given polymorphism within autosomal genes. Again, predicting the phenotype of heterozygous individuals will be an important issue in undertaking association studies. To return to the example of CCR5Δ32, individuals heterozygous for

this deletion would be expected to have half the level of expression of CCR5 at the cell surface compared with individuals homozygous for the wild type form of the receptor. Even here, where the functional consequences of the deletion are clear cut, the phenotype of heterozygotes may be difficult to predict: would one expect an individual with half the level of CCR5 expression on macrophages to be resistant to HIV entry? If it is unclear what the phenotype of heterozygous individuals is likely to be from functional data (which will be the case for the vast majority of SNPs currently being described), then association studies will have to look both at the carriage of the allele and also separately at the homozygous groups. This again increases the number of comparisons being made and requires adjustments to be made to the statistical analysis.

A further point which seems obvious but which is often forgotten is that the size of the population required for a study will depend upon the allelic frequency of the polymorphism (or combination of polymorphisms) under consideration. This may be known from published data but sometimes it is uncertain at the time a study is planned. For example, several years ago we undertook a study looking at the potential contribution of a SNP which resulted in an amino acid substitution in the IL-9 gene to variability in IgE levels in a random population. Reasonable functional data predicted that we might expect to see lower IgE levels in those homozygous for this polymorphism. We studied over 600 individuals but, because the allelic frequency of the polymorphism in this population was low, we only had eight individuals homozygous for the polymorphism. As a result, while we could be reasonably certain that individuals heterozygous for this IL-9 polymorphism do not have lower IgE levels, we were unable to determine the potential contribution of the polymorphism in those homozygous at this locus because of inadequate numbers despite the apparently reasonable size of the study.

The next major problem in candidate gene association studies is the use of inappropriate control populations. For example, a study of asthma candidate gene polymorphisms which used anonymous blood donors as a control would be flawed because a significant number of the control population may well have mild asthma. This is perhaps less of a problem with rarer diseases—for example, a study on idiopathic pulmonary fibrosis which used age matched blood donors would be unlikely to have any cases among the control population—but, ideally, the control population should be phenotyped in the same way as the study population. For common diseases such as asthma this is relatively straightforward: the use of large cohorts collected in epidemiological studies is one obvious approach.

It is also important to prevent the occurrence of population stratification. The best example of this is the use of groups of different racial origin in the control and affected populations. The prevalence of polymorphisms within different racial groups is often markedly different and hence false positive (and negative) results may result by the selective inclusion of a small number of individuals from a different racial group in the control or affected populations. Most studies now limit their analysis to groups of the same racial origin, but even this is potentially open to criticism. For example, returning to the example of the CCR5Δ32 mutation, there is a marked difference in the prevalence of this mutation between white populations in northern and southern Europe. This can obviously cause problems where large association studies have been performed involving pooled samples from a number of centres.

The problem of multiple testing with reference to genotypes versus alleles has already been discussed, but the main problem with multiple testing is the use of multiple phenotypes. The best way to avoid this is to define in advance the primary and secondary end points to be studied. This does not preclude the reporting of an association with a different end point from those which have been initially defined, but such an association should be reported as a chance finding in the study which requires verification in a separate population. One difficulty in this area is the repeated use of the same data set for association studies without appropriate correction for multiple testing. For example, an investigator might collect 500 asthmatic and 500 control individuals and look for an association in that population with a given gene polymorphism and report on that association. If he or she were then to go back to look at 10 other genes in that population, it would be inappropriate to report each of those association studies as a separate study without making allowance for the multiple testing. This can easily be spotted by a reviewer of the study if data on all of the polymorphisms are included in a single paper, but unfortunately this is not always the case. Investigators should be honest in reporting the number of association studies carried out in a given population and make allowance for this either in the analysis or in the discussion. In this respect, gene association studies are no different from epidemiological studies in the ease in which data mining approaches can be used.

Notwithstanding the above concerns, candidate gene studies will be important in establishing the basis for interindividual variability in disease susceptibility, severity, and response to treatment. The key issue is that investigators must consider the above factors in their study design before performing an association study and, most importantly, they must ensure that the size of the proposed study is adequate to address the question being asked. In practice, most polymorphisms will have relatively small effects rather than be all or none disease causing mutations, hence the size of populations required to study their contribution will be large. Most studies examining the contribution of polymorphisms in candidate genes in polygenic diseases such as asthma or chronic obstructive pulmonary disease will need at least 500 individuals in each group and, where several end points or multiple polymorphisms are being considered, numbers will often be over 1000. Given that it is difficult for single investigators to generate populations of this size, the need for collaboration between groups is clear.

## ACKNOWLEDGEMENT

Work in the author's laboratory on asthma genetics has been funded by the NAC, MRC and the Wellcome Trust.

*Thorax* 2002;**57**:377–378

**Author's affiliation**

**I Hall,** Division of Therapeutics, Queen's Medical Centre, University Hospital, Nottingham NG7 2UH, UK

Correspondence to: Professor I Hall, Division of Therapeutics, Queen's Medical Centre, University Hospital, Nottingham NG7 2UH, UK; ian.hall@nottingham.ac.uk

## REFERENCES

1 **International Human Genome Sequencing Consortium**. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
2 **Sachidanandam R**, Weissman D, Schmidt SC, *et al*. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;**409**:928–33.
3 **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**:720–31.
4 **Dean M**, Carrington M, Winkler C, *et al*. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science* 1996;**273**:1856–62.
5 **Samson M**, Libert F, Doranz BJ, *et al*. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;**382**:722–6.