Alcohol & Alcoholism Vol. 36, No. 2, pp. 99–103, 2001

INVITED REVIEW

GENETIC ASSOCIATION STUDIES OF ALCOHOLISM — PROBLEMS WITH THE CANDIDATE GENE APPROACH

PAUL R. BUCKLAND

Department of Psychological Medicine, University of Wales College of Medicine, Heath Park, Cardiff CF14 4XN, UK

(*Received* 17 *April* 2000; *and in revised form* 21 *September* 2000)

Abstract — In recent years, progress has been made in the identification of causative factors in most single gene disorders and those with genes of major effect. In comparison, no genes contributing to a complex disorder have been unambiguously identified. A number of reasons for this have been previously presented in theoretical papers. Alcoholism is such a complex illness and genetic studies into its underlying genetic causes have suffered from lack of power due to small subject numbers, poor selection of control subjects, and over-emphasis on markers with low prior probability of involvement.

INTRODUCTION

In the last 10 years, there has been a growing list of genetic linkage and association studies of alcoholism and the diseases resulting from alcoholism. Unfortunately, not only have these studies been inconsistent and inconclusive, but also many studies described in submitted manuscripts to medical and scientific journals are poorly designed. I describe below some of the problems with such studies and their design, and factors which should be taken into account before such studies are undertaken, or indeed funded.

It is well established that alcoholism has a genetic component, but it is a complex syndrome, possibly one of the most complex known. There is a considerable body of literature on the issues and strategies concerning genetic studies to find genes contributing to complex illnesses (Owen *et al*., 1997, 2000) including alcoholism and its associated illnesses (Comings, 1998; Conneally and Sparkes, 1998; Goldman *et al*., 1998; Hill, 1998; Schork and Schork, 1998; Noble, 1998). Finding genes predisposing to complex phenotypes is far from simple, and is likely to require large multi-disciplinary research groups accumulating hundreds, and probably thousands of DNA samples from rigorously phenotyped subjects. In addition, with the association design, in order to have a reasonable *a priori* chance of analysing genuine susceptibility genes (see section below), it will usually be necessary to have access to fairly high (and expensive) laboratory technologies and expert statistical guidance. In alcoholism, it is reasonable to expect that the environmental effects will be more problematic than in other complex phenotypes such as schizophrenia or bipolar disorder (Hill, 1998). For example, the development of alcoholism must be preceded by exposure to alcohol; this is rare in some social groups, whilst common in others, and is affected by free will and prior knowledge of the effects of alcohol. In comparison, the only known environmental effectors of schizophrenia are place and season of birth (Mortensen *et al*., 1999) neither of which can be controlled by the subject. It is likely, therefore, to be a more difficult task finding genes predisposing to alcoholism than to psychosis. In research on alcoholism, one would expect to see larger data samples, better defined patient phenotypes and control groups, and more sophisticated analysis than is reported in studies of psychosis. Unfortunately, the opposite is frequently the case.

In addition to the obvious need to refine the ‘alcoholism’ phenotype, particular attention needs to be given to two aspects of alcohol association studies: the phenotype of the controls and the size of the patient sample.

CONTROLS

One of the principal problems with genetic studies of alcoholism is the ascertainment of suitable controls. It is axiomatic to science that to determine the effect of one variable the researcher must keep all other variables either constant or randomized. The perfect control for propensity to dependence on alcohol (as a for example phenotype) is a group matched in every way possible to the affected group. Thus, the group should be matched at least for age, sex, weight, ethnic and social background, current major life events (and other environmental variables) but above all else, alcohol intake. The latter is important, because it is likely that the major environmental influence on alcohol dependence is the drinking of alcohol. Over the entire research sample, this must be either kept constant or be randomized. As all alcohol-dependent subjects by definition have drunk large quantities of alcohol, then this factor cannot be randomized. It clearly follows that control subjects then must have had (at some stage in their lives) similar exposure to alcohol with respect to intake as the affected group, and, what is more, this should ideally be in a social setting where it was similarly acceptable to become intoxicated. Such a control group is not easy to come by (however, might I suggest university alumni; British students at least are likely to conform to the above criteria).

Clearly, the optimal control group will depend upon the particular part of the alcoholism phenotype under investigation. For example, if one is studying the propensity for alcohol abuse, a measure of alcohol exposure in the control group, as above, is not necessary, as genes which contribute to addictive behaviour or sensation-seeking (or other psychological constructs) or illnesses such as depression, may lead a person to seek out alcohol. However, in this case, a lack of specificity is likely to be encountered and it is generally assumed that genes for addictive behaviour may moderate either a psychological or biological drive unrelated to the specific effects of alcohol. However, instead of selecting the most appropriate groups, commonly, controls are taken from a different study, from a group of patients at a clinic or blood donors, including donations from hospital staff or unaffected family members.

Each of these presents a problem. The first two can suffer, not only from population stratification (one subgroup being genetically different from another), but also the gene under investigation may be involved in the syndrome of the original study or illness and thus not be a control group at all. Blood donations (from the general population or hospital staff) are voluntary and require certain characteristics in the donor, for example altruism or financial hardship, which may be relevant to conditions such as alcoholism. In the UK, blood donations are frequently taken in a batch from one location (factory or hospital) and they are not screened for most illnesses including alcoholism (although ongoing treatment for a condition leads to exclusion, and so many chronic illnesses can be discounted).

Family-based association studies can have advantages where there is population stratification (Owen *et al*., 1997) and this form of analysis may also control for many putative environmental effects (e.g. social tolerance of alcohol abuse) mentioned above. However, these studies may have reduced power, particularly if the families studied have high frequencies of the susceptibility allele. Nevertheless, it is likely that they are less frequently used in studies of alcoholism due to the considerable work involved in contracting, characterizing and taking blood from patients' relatives, compared with unrelated controls.

Some of these problems are illustrated in a meta-analysis of association studies between a polymorphism close to the dopamine D_2 receptor (DRD2) and alcoholism (Noble, 1998). This latter author reported a highly significant result based upon 15 studies which used a polymorphism in the 3′ flanking region of DRD2. However, a very large analysis of sib-pairs using this same polymorphism and also another functional exonic polymorphism which severely impairs the function of the D_2 receptor, found no association (Goldman *et al*., 1997, 1998).

The studies above do not necessarily repudiate each other, because they were carried out in different racial groups which may have quite different environmental and genetic risk factors that interact with the DRD2 'risk' allele. Also differences in the DRD2 allele frequencies may lead to a different power to detect the effect. In this case, the population studied by Goldman *et al*. (1997) had a much higher *A1* allele frequency than the populations referred to by Noble (1998). However, those studies reviewed by Noble (1998) and carried out on Caucasians appear to show considerable stratification. The frequency of the *A1* allele in the control samples varies from 6.0 to 21.0%, a 3.5-fold variation. Where no significant association between the *A1* allele and alcoholism was found, the control *A1* allele frequency ranged from 15.0 to 21% (7 studies), whereas when a significant association was found in six studies, the *A1* allele frequency ranged from 6 to 11.1% and only two were above this range (17.2 and 18.5%). Noble (1998) attributed this apparent bimodal distribution to the inclusion of alcoholics in some of the controls (even though the latter two positive studies with the high control *A1* allele frequency did use screened controls). If the differences in allele frequency between the control groups were to be explained by non-exclusion of alcoholics, most of the subjects in the non-screened control samples would have had to have been alcoholics. Of course this would be possible if controls were paid blood donors.

Any association study tests the null hypothesis that the frequency of a particular allele (or genotype) occurs with the same frequency in patients and controls. Where more than one study has taken samples from the same pool, the data may be combined additively in a meta-analysis. However, where the pools are different, as seems to be the case in the studies described by Noble (1998), this is not valid and may lead to Simpson's paradox, namely that: it is not necessarily true that averaging the averages of different populations gives the average of the combined population. Instead, some form of weighted paired analysis is required, such as that described by Woolf (1955).

The above comments are not meant as an attack on any of the authors mentioned, nor as an explicit opinion that the *A1* allele of DRD2 is not a risk factor involved in alcoholism. I merely wish to point out that the phenotypes and genotypes of the controls are as important as those of the patients, especially when studies are compared and meta-analyses are carried out.

As to the *Taq1A* (*A1*) allele itself, a commonly quoted axiom is that extraordinary claims require extraordinary evidence. The *Taq1A* polymorphism lies over 10 kb downstream of the termination codon of the DRD2 gene and 250 kb away from the start of DRD2 transcription. It was initially characterized for use in family-based linkage studies, which can detect effects over long chromosomal distances. It seems unlikely to be in linkage disequilibrium with another polymorphism which affects transcription or to do so itself. It is a very dim light in an extremely long street.

POWER

Statistics should be an aid to logic, not a replacement for logic. All statistical analyses make assumptions which in any given data set may not be true. Thus, statistics can never prove or disprove a hypothesis, but can indicate how likely it is to be true if the assumptions are correct. An example of statistics taking priority over logic is shown by a letter commenting on the study of Goldman *et al*. (1997). Vanyukov (1999) suggested that the frequency of Cys/Cys homozygotes varies significantly between substance abusers (0.053) and non-abusers (0.022). These figures look impressive until one realizes that they actually represent nine and six individuals respectively, and any statistical analysis of these numbers is ludicrous.

The common practice of counting allele frequency may also distort the results of an association, as there are twice as many alleles as people and each person only has one chance of being an alcoholic. Some commentators consider that *P*-values are therefore artificially lowered and significance increased by this 'double counting' (Sasieni, 1997). More importantly, when alleles, rather than genotypes are analysed, both the odds ratio and the χ^2-test are inappropriate, unless the Hardy–Weinberg equilibrium is maintained. If it is not, there will be an increase in false positives (Sasieni, 1997).

The power of any study to detect the genetic effect is all too often overlooked in association studies. A sample power analysis for schizophrenia association studies has been described (which assumes that the actual polymorphism used contributes to the phenotype or is in complete linkage disequilibrium with such a polymorphism; Owen *et al*., 1997). In simple terms, three pieces of information are required in order to calculate how many samples are required for an association study: the desired *P*-value, the size of the effect expected, and the frequency of the alleles to be used. The size of the effect is often referred to as the relative risk, and smaller relative risks generally require larger sample sizes in order to be detected. If a genetic polymorphism used as a marker in an association study has two alleles present at equal frequency in the population under study then 25% of individuals will be homozygous for allele *A*, 25% homozygous for allele *B*, and 50% heterozygous (Hardy–Weinberg equilibrium). As the minor allele becomes less common, the number of samples that are required to achieve the same power changes. If the relative risk is 2 or below, any change from a 50% frequency lowers the power (Risch and Merikangas, 1996). The attributable risk (Owen *et al*., 1997) is the proportion of cases which would disappear if the illness causing polymorphism ceased to exist. It is a function of the relative risk and the allele frequency, and therefore need not be considered in addition to them.

A *P*-value of 0.05 means that the result obtained would have occurred by chance 5% of the time. Therefore, if 20 independent tests are carried out, it is likely that one false positive will be obtained by chance alone. Published results have been shown to conform to the expected pattern in a survey by Terwilliger and Weiss (1998). What a *P*-value of 0.05 does not mean is that there is 95% confidence that the result is a true positive.

As in many areas of research, genetic association analysis usually requires that the researcher carries out many independent tests. If one in 20 of those tests gives false positive results by chance, then the problem is one of telling the false, from the true, positives. This problem is often described as being one of multiple testing: it is not. The Bonferroni correction, whereby the threshold for statistical significance is multiplied by the number of tests (Owen *et al*., 1997), is frequently applied. However, this multiple testing fallacy and the use of ‘corrections’ has been debunked previously (Perneger, 1998). The fallacy originates due to the misconception that *P*-values indicate the power of an analysis: they do not. The multiple testing problem is only relevant to resampling of the same pool; take for example the statement ‘eight out of ten dogs prefer …’. In this case, the dog food manufacturers have tested a number of different groups of 10 dogs but only reported the results from one such sampling. They have not quoted an average. No matter how horrible the food, eventually they will find a group of 10 dogs of which eight like it. This is an example of multiple testing of the same hypothesis. When the same hypothesis such as ‘the dopamine D_2 receptor is associated with alcoholism’ is repeatedly tested using several independent polymorphisms (ones not in linkage disequilibrium) then the Bonferroni correction may be applicable (Boehringer *et al*., 2000). It is not clear to me why, but this argument is considered to be controversial.

What must be remembered by the non-statistician is that *P*-values do not directly indicate how likely it is that any hypothesis being tested is true or false and the ‘significance level’ of 0.05 has to all intents and purposes been picked out of a hat. More informative statistical aids to logic in this area are available, such as odds ratios, conditional probabilities, Bayes’ theorem, and likelihoods.

When the term ‘multiple testing’ occurs in biological research or health issues, what the scientist usually wants to know is if a positive result has been obtained, how likely is it that it is a true positive; i.e. the likelihood ratio (for a detailed explanation see: http://mathworld.wolfram.com/LikelihoodRatio.html). This is simply the chance of finding a true positive divided by the chance of finding a false positive (or sensitivity divided by 1 — specificity). In practice, it is likely to be impossible to calculate or even estimate the likelihood ratio as the sensitivity is unknown. If the probability of finding a true positive is quite high, then this problem can reasonably safely be dealt with by repetition of experiments. However, in research where true positives are likely to be rare, it is essential that some attempts be made to estimate the ratio of true to false positives.

If a single association test is carried out and the resulting *P*-value is 0.05, we know that this result would have occurred by chance one in every 20 experiments or 5% of the time. How often would we expect to obtain a true positive result? If we assume that: (1) all genes have an equal chance of being associated with alcoholism; (2) there are 20 genes genuinely associated with alcoholism; (3) there are 80 000 different genes; (4) the role of each gene can be fully tested using a single polymorphism. Then we would expect to find a true positive result in one in every 4000 experiments or 0.025% of the time. Therefore any result with the *P*-value of 0.05 has a 99.5% likelihood of having occurred by chance. If the *P*-value obtained was 0.005 (0.5% of genes would give this value by chance i.e. 400 genes) there is still a 95% likelihood that this is a false positive. Extending this calculation, a *P*-value of 0.00025 would mean that there was a 50% likelihood that it was a true positive and to get a 95% likelihood of a true positive, a *P*-value of 0.0000125 is required.

If we now look at our assumptions, (4) is clearly wrong. The number of polymorphisms required to screen an entire gene is unknown, as it will vary considerably between genes and between populations, but will almost always be greater than one. The question here is, if there is a polymorphism in a gene which genuinely contributes to alcoholism, but we pick a different polymorphism in the same gene which does not contribute to alcoholism, how likely is it that an association study between the latter polymorphism and alcoholism will give a positive result by virtue of linkage disequilibrium. The answer is that there is no way of knowing. Two polymorphisms 10 bases apart may not be in linkage disequilibrium, whereas those many thousands of bases apart may be. Kruglyak (1999) has estimated that on average, useful degrees of marker–marker linkage disequilibrium do not extend beyond 3000 bases. Using a similar figure to this, Risch and Merikangas (1996) have estimated that to get a 95% probability of no false positives on a genome wide association study, a *P*-value of 5×10^{-8} is required.

There may be ways out of the dilemma of unreachable *P*-values. We can look at assumption (1) above, from a Bayesian standpoint and use prior probability. If we only study functional alleles in genes which we estimate are (hypothetically) 100 times more likely to be involved in alcoholism than other genes, we can increase the *P*-value required by 100 times. Clearly such an estimate cannot be truly quantitative, but it does illustrate the reason for studying good candidate genetic polymorphisms for alcoholism, rather than polymorphisms which have no merit other than that they are well known.

It has been suggested that the whole set of patient and control DNAs can be divided into two groups; an initial association study is done on one group, and if a *P*-value lower than a chosen threshold is found, the study is repeated independently on the second sample (Owen *et al*., 1997). If the results concur, then there is a far better chance of the results being correct. However, I can think of no rationale for doing this other than saving work. There is also the risk of missing a true positive if the first group is not large enough.

GENERAL CONCLUSIONS AND COMMENTS

In summary, without extremely large samples ($n = 1000$), researchers should restrict themselves to hypotheses with at least a modest probability of being correct. They need to determine that their experiments have the power to detect the effect in question (to avoid false negatives) and all positive results should be assumed to be false until independently replicated. When the power of a study is lower than is ideally required, this should be acknowledged by the authors, who should refrain from drawing unsubstantiated conclusions and instead present their data as part of an ongoing global study.

Whilst one conclusion from the above is that only well resourced research groups should try and find genes for alcoholism, smaller groups can make a useful contribution, but careful consideration needs to be given to the design and reporting of the work.

For any gene which does contribute to alcoholism, its potential attributable and relative risks are unknown and can only be guessed at. Therefore, any minimum sample size is, on a scientific basis, entirely arbitrary even before the question of *P*-values is considered. The minimum number should therefore be set by practical parameters. The power analysis of Owen *et al*. (1997) shows that several hundred subjects are required to achieve an 80% chance of detecting an association with a relative risk of less than 2 at the $P = 0.05$ significance level. For a lower *P*-value, much larger numbers are required.

What must also be considered in the light of the above comments is that, although I have concentrated on the problem of false positives, many of the factors related both to the power and to the selection of subjects and controls will increase the probability of false negatives. The latter type of error is possibly more important, as a negative association study is far less likely to be replicated than is a positive one.

One the one hand, it is unreasonable to expect every study to involve several thousand subjects and controls. On the other, a group so small that an effect of the size to be expected cannot possibly be detected above the background noise of chance findings, is scientifically pointless. It has been suggested that the minimum number of patient DNA samples that can be usefully analysed is 100, in addition to 100 matched control samples (Conneally and Sparkes, 1998). This seems appropriate if very common polymorphisms are to be studied. For less common polymorphisms, even more samples are generally required. However, it must be remembered that any analysis of the data from such a small sample is most probably virtually meaningless, and the results are only of use in a meta-analysis. Therefore, both patients and controls need to conform to an internationally agreed standard to be included in such an analysis. Here, the suggestion by Noble (1998) that the comparison groups should be mild alcoholics vs severe alcoholics seems to be one possibility, as many of the problems with controls are attenuated.

There are three stages to any genetic study: first collecting the samples and preparing the DNA, second carrying out mutation analysis to detect novel polymorphisms, and third analysing the DNA for association between those polymorphisms and an illness. The first of these is expensive and time-consuming, and it may take many years to collect samples depending on many factors including the complexity of the diagnostic work up. Similarly, screening an entire gene for polymorphisms can take many months depending on several factors such as number of exons or GC content of certain regions, etc. However, the analysis of DNA using modern methods typically allows the genotyping for one polymorphism on 400 samples in 2 days if done manually, and considerably less time than this if robotics are used. Alternatively, four polymorphisms can be genotyped on 100 samples in 2 days. It is clear therefore, that a report of an association study for one polymorphism on 50 patients and 50 controls is by these standards a very small amount of work and not a publishable unit, unless the results make a significant contribution to knowledge, which is unlikely given the analysis above. Whilst it might be argued that allowances should be given to less well resourced laboratories to promote the development of research, it must be recognized that the ‘minimum publishable unit’ must change with time; for example in the early 1980s the sequencing of a few hundred bases of a gene would be enough for a publication; clearly this is no longer the case.

A publishable unit of work should include the following: either the first reporting of the DNA sample collection and its characterization, or mutation analysis to detect novel polymorphisms, in addition to a genetic analysis, such as an association study (when reporting the DNA sample for the first time a single analysis may constitute enough work to justify the publication, although the collectors of the samples should expect to be the primary authors), or for subsequent analyses using previously published samples and polymorphisms, several thousand individual analyses (number of DNA samples multiplied by the number of polymorphisms) unless an important positive finding is included.

A researcher planning a genetic association study for alcoholism or an associated illness needs to have the following: (1) a good rationale for studying not only the gene in question, but the specific polymorphism; (2) enough subjects and controls for meaningful analyses; (3) a valid set of control DNA samples well matched to the subject samples to eliminate as many of the environmental effects leading to alcoholism as possible.

In addition, it would be helpful if those involved in alcohol research could agree on standards for the collection of patient and control samples to enable meaningful meta-analyses and to agree on a minimum number of subjects required for publication. To take into account the difficulty in obtaining large samples, a new publishing format may be considered appropriate for smaller studies. Perhaps, studies carried out on previously reported samples should take up no more than half a page. All that is required in addition to the raw data are references to the DNA samples, polymorphisms, and previous related studies which should describe the rationale and procedures.

In 1993, Kidd discussed the statistical problems with association studies and posed the question '… are association studies even worth doing'? (Kidd, 1993). Since then, many commentaries on this subject have been published, including ones in psychiatry-related journals. Paterson (1997) suggested that journals should refuse to accept any reports of case control genetic association studies of complex traits at all!

Ultimately, the question is one of efficient use of resources. Since the studies described above were published (Noble, 1998) there have been at least 10 further association studies between the DRD2 gene and alcoholism with the majority showing no association. It is not clear how much more effort and public money will be directed towards this question before any widely accepted conclusion can be drawn.

REFERENCES

Boehringer, S., Epplen, J. T. and Krawczak, M. (2000) Genetic association studies of bronchial asthma — a need for Bonferroni correction? *Human Genetics* **107**, 197.

Comings, D. E. (1998) Why different rules are required for polygenic inheritance: lessons from studies of the DRD2 gene. *Alcohol* **16**, 61–70.

Conneally, P. M. and Sparkes, R. S. (1998) Molecular genetics of alcoholism and other addiction/compulsive disorders. *Alcohol* **16**, 85–91.

Goldman, D., Urbanek, M., Guenther, D., Robin, R. and Long, J. C. (1997) Linkage and association of a functional DRD2 variant [Ser311Cys] and DRD2 markers to alcoholism, substance abuse and schizophrenia in southwestern American Indians. *American Journal of Medical Genetics* **74**, 386–394.

Goldman, D., Urbanek, M., Guenther, D., Robin, R. and Long, J. C. (1998) A functionally deficient DRD2 variant [Ser311Cys] is not linked to alcoholism and substance abuse. *Alcohol* **16**, 47–52.

Hill, S. Y. (1998) Alternative strategies for uncovering genes contributing to alcoholism risk: unpredictable findings in a genetic wonderland. *Alcohol* **16**, 53–59.

Kidd, K. K. (1993) Associations of disease with genetic markers: Déjà vu all over again. *American Journal of Medical Genetics* **48**, 71–73.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* **22**, 139–144.

Mortensen, P. B., Pedersen, C. B., Westergaard, T., Wohlfahrt, J., Ewald, H., Mors, O., Andersen, P. K. and Melbye, M. (1999) Effects of family history and place and season of birth on the risk of schizophrenia. *New England Journal of Medicine* **340**, 603–608.

Noble, E. P. (1998) The D2 dopamine receptor gene: a review of association studies in alcoholism and phenotypes. *Alcohol* **16**, 33–45.

Owen, M. J., Holmans, P. and McGuffin, P. (1997) Association studies in psychiatric genetics. *Molecular Psychiatry* **2**, 270–273.

Owen, M. J., Cardno, A. G. and O'Donovan, M. C. (2000) Psychiatric genetics: back to the future. *Molecular Psychiatry* **5**, 22–31.

Paterson, A. D. (1997) Case-control association studies in complex traits — the end of an era? *Molecular Psychiatry* **2**, 277–278.

Perneger, T. V. (1998) What's wrong with Bonferroni adjustments. *British Medical Journal* **316**, 1236–1238.

Risch, N. N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Sasieni, P. D. (1997) From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.

Schork, N. J. and Schork, C. M. (1998) Issues and strategies in the genetic analysis of alcoholism and related addictive behaviours. *Alcohol* **16**, 71–83.

Terwilliger, J. D. and Weiss, K. M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9**, 578–594.

Vanyukov, M. M. (1999) An association between a functional polymorphism at the DRD2 gene and the liability to substance abuse. *American Journal of Medical Genetics* **88**, 446–447.

Woolf, B. (1955) On estimating the relationship between blood groups and disease. *Annals of Human Genetics* **19**, 251–253.