Epilepsia, **47**(10):1617–1621, 2006
Blackwell Publishing, Inc.
© 2006 International League Against Epilepsy

Editorial Commentary

Can We Increase the Likelihood of Success for Future Association Studies in Epilepsy?

*Martina Durner, *Prakash Gorroochurn, †Carla Marini, and †Renzo Guerrini

**Division of Statistical Genetics, Columbia University, New York, New York, U.S.A., †Epilepsy, Neurophysiology and Neurogenetics Unit, Department of Child Neurology and Psychiatry, University of Pisa and Research Institute Stella Maris Foundation, Pisa, Italy*

Recently, association studies have become a popular tool to elucidate the genetic basis of diseases that show a high familial recurrence rate but probably are not caused by one single gene. These diseases, like many of the common forms of idiopathic generalized epilepsy, are thought to be caused by several interacting genes, with perhaps also an environmental component to the disease mechanisms. Each of these genes is supposed to have only a small effect and is neither necessary nor sufficient to cause epilepsy by itself.

Association studies are considered to be a better tool to detect those genes of small effect than, for example, linkage studies (Risch and Merikangas, 1996). Genetic association is observed at a population level by assessing correlations between genetic variants and the epilepsy phenotype reflecting the original genotype (haplotype/ genetic makeup) of a single founder that has been altered by historical recombinations. However, proof that association indeed exists would need replication of the original findings in independent samples, and ultimately, demonstration that the associated allele(s) can contribute to the disease phenotype by altering biological functions (Colhoun et al., 2003).

We expect the number of association studies in epilepsy to increase in the future because of the relative ease of collecting subjects (only patients and healthy controls are needed compared to whole families in linkage studies) and determining genetic variants (high through-put facilities can process thousands of genotypes per day at ever decreasing costs).

The expanding use of this association approach raises several issues, some of them exemplified by the manuscript by Barratt and colleagues published in this issue of *Epilepsia*. Sander et al. (2000) reported an association of idiopathic absence epilepsy (IAE) with a functional variant in the μ-opoid receptor (OPRM) ($p = 0.02$). In their replication study, Barratt and colleagues found no difference in the frequency of this variant between their cases with either IAE or idiopathic generalized epilepsy and their controls. But they also noted a key difference between their study and that of Sander et al.: the frequency of this variant was statistically highly significantly different between *control* groups ($p = 0.0001$)!

How can we explain this observation? There are several scenarios to consider depending on whether we assume that there truly is or is not an association with OPRM and epilepsy. Under the assumption of a true association, the failure of Barratt and colleagues to replicate this finding can be caused by either low power of the sample, population stratification in their sample, or inadequate control group. Barratt and colleagues addressed the issue of power in their manuscript and showed that their sample is large enough to detect an association if there is one. Population stratification (PS) refers to the situation when the sample is drawn from two different populations, each differing in the frequency of the disease and the marker alleles. PS is very unlikely in this case, because in order to mask the assumed association in one population, the other population would have to have a negative association with the same allele of OPRM (i.e., in one population this allele is contributing to the disease whereas in the other population it is protective). The higher frequency of the variant in the control sample of Barratt and colleagues could be caused by the existence of epilepsy cases among the control sample. This is also very unlikely, because Barratt and colleagues report that all controls were screened for epilepsy.

Under the assumption of no association, we have to declare Sander et al.'s (2000) result a false positive. One major concern of spurious association results is

Accepted July 24, 2006.
Address correspondence and reprint requests to Dr. Martina Durner, Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W. 168th Street, New York, NY 10032, U.S.A. E-mail: md2050@columbia.edu
doi: 10.1111/j.1528-1167.2006.00842.x

population stratification. Sander et al. (2000) report only that all their patients and controls are of German descent. It is possible that their sample contains some hidden population structure that is not obvious from the geographical area from which they were drawn. However, in a subsequent paper published by the same group (Lorenz et al., 2006), the authors performed a structured analysis in a much larger sample and found roughly equal frequency distributions at neutral loci among their patients and controls. It is therefore likely that the sample used in the initial analysis (2000) is contained in the latter analysis (2006), and that even if there is cryptic population substructure in this sample, the effect is probably minimal.

Another, albeit unlikely, explanation of the low frequency of the genetic variant in OPRM in the control sample of Sander et al. is natural selection *against* this variant in controls. This leaves us with the most likely explanation for the observed constellation—that random variation of the frequency of the genetic variant in the controls has led to the observed association.

When we perform a statistical test, we always calculate the p-value, or the probability of obtaining a result as extreme as (or more extreme than) that observed, given that chance alone is present. Using a nominal significance level of 0.05 this probability is 5%, indicating that under the null hypothesis, 1 of 20 tests on average is a false positive. However, when we are using the same data set to test for association with various candidate genes or chromosomal regions, we are performing more than one test. In every region, usually several markers with a variable number of alleles are tested. Each allele is tested for correlation with the disease, thus representing one test. Multiple testing therefore becomes an issue, and the probability of obtaining a false positive result increases. The most commonly used method to correct for this increase in type I error is the Bonferroni method, in which the nominal p-value (i.e., the p-value obtained for one test) is multiplied by the number of tests performed. The Bonferroni method is sometimes conservative and overcorrecting because often the markers are not truly independent. Another concern is that we are looking for genes with small effect, and by applying too stringent significance criteria we might miss a true association.

Skepticism about the correctness of many positive findings, derived from the observation that consistent replication is only obtained for a very small minority of association studies (Hirschhorn and Altshuler, 2002; Healy, 2006), has led some journals to produce guidelines for genetic association studies of complex disorders (Editorial, 1999; Bird et al, 2001; Little et al, 2002; Cardon and Bell, 2001) or to openly discourage submission, unless key biologic insights are disclosed (Marks, 2004). Guidelines for genetic association studies in epilepsy were published in *Epilepsia* in 2004 (Tan et al., 2004).

A key issue in increasing confidence in reported associations is replication and biological plausibility. However, nonsignificant or negative replication studies, when designed properly, will be also helpful to identify false positive results. As the study by Barratt and colleagues showed, such studies can also highlight the importance of rigorous design and evaluation strategies for future association studies. Certain basic precautionary steps should be taken so that future positive findings are viewed with more credence, and not merely as false positives or artifacts:.

TESTING (AND CORRECTING) FOR POPULATION STRATIFICATION

There exist several methods that either avoid PS, or test and correct for it. The Transmission-Disequilibrium Test (TDT) was devised as a test for association that is not affected by PS (Spielman et al., 1993). It measures the preferential transmission of an associated allele from parents to offspring and thus requires “trios,” that is, cases and their respective parents. Such samples may be harder to collect than the standard samples for association studies; further, the TDT turned out to be not very powerful. A standard case–control design has been preferred therefore by many researchers (Healy, 2006). However, cases and controls should be matched for ethnicity, and preferably should come from the same source population (or geographical area). Although this caution might reduce confounding by PS, it is always possible for cryptic substructure to exist in the study population. A simple chi-square test for Hardy–Weinberg equilibrium (HWE) in the controls can detect substructure, but this test is well known to have poor power unless very large samples ($\sim$500) are used (Halliburton, 2003). If only cryptic substructure is present, this test would be hopelessly underpowered. Thus, failure to detect deviations from HWE in controls does not discard PS of the magnitude that can adversely affect association studies.

Pritchard and Rosenberg (1999) proposed the use of a panel of markers unlinked to, and also not in linkage disequilibrium (LD) with, the test locus, known as null (or neutral) markers. These markers can be used to test for PS in the sample of cases and controls. This test is based on the fact that the sum of the chi-squares at these neutral loci has a chi-square distribution under the hypothesis of no PS. Any deviation from this distribution therefore has to be considered as an indication for PS. More recently, a related test has been proposed, which has the flexibility of being able to test either for any PS or only for PS that is practically important (Gorroochurn et al., 2006a).

PS will lead to an inflation of the chi-square test for association and thus to more false positives than the nominal significance level indicates. Several methods have been recently proposed to correct for this inflation. Using the

information from the neutral loci, the chi-square statistic can be adjusted to reflect this inflation at the test locus (also called Genomic Controls (GC); Devlin and Roeder, 1999; Devlin et al., 2001). Another method to approach the problem of PS is structured association (SA; Pritchard et al., 2000a; 2000b). SA applies a Bayesian clustering method to a panel of neutral loci to infer the number of subpopulations and the subpopulation membership of subjects in a sample. Association between marker and disease is then tested in each of these identified subpopulations. Gorroochurn et al. (2006b) proposed a third method, delta-centralization (DC), which uses a better approximation of the test statistic than GC and avoids the computational intensiveness of SA. It is easy to apply and furthermore requires a relatively low number of neutral loci (about 20) when the allele frequencies at these loci are matched to those at the test locus.

We conclude, therefore, that with little increase in time and money, positive association findings can be made more robust and should become the "gold standard" for future published results.

CORRECTING FOR MULTIPLE TESTING

As stated above, if several successive tests are performed on the same data set, the overall (experiment-wise) type I error (i.e., falsely rejecting the null hypothesis of no association) increases. A reported p-value of 0.02 for one experiment is associated with a different significance level if only one experiment is performed or if it is part of a series of tests. In the case of multiple testing, one should either correct the p-value or at least report all the tests performed so the reader can make an informed judgment and interpretation of the finding.

Traditionally, the problem of multiple testing has been addressed through a Bonferroni correction, which consists of performing each of the L tests at an α/L level (where α is the experiment-wise error). In genetic association studies however, the number of tests performed becomes quickly prohibitively large, thus making this procedure very conservative and the study underpowered to detect small but true effects. For this situation, Benjamini and Hochberg (1995) introduced the False Discovery Rate (FDR)—the expected ratio of erroneous rejections to the number of rejected hypotheses—as an appropriate error rate to control.

SAMPLE SIZE: PROS AND CONS

Pros: Under the "common disease, common variant" hypothesis, sample size is of concern. Here we hypothesize that a genetic variant is neither necessary nor sufficient to cause disease, that it is relatively frequent, and that it is also found in controls albeit less often. Each of these factors will contribute to requiring a large sample size to detect a small effect.

But once an association is found, it is important to replicate this finding (to confirm or refute it), especially since so many association reports have turned out to be spurious (e.g., Healy, 2006). Power is specifically an issue in replication studies that fail to replicate a reported association. The failure to demonstrate an existing association might be because the study is not sufficiently powered to detect a small effect size. In addition, it has been shown that the effect size is often overestimated in initial studies and tends to be lower in replication studies, thus requiring an even larger sample size than in the initial positive study.

But even when equal effect size is assumed in both initial and replication studies, Greenwald et al. (1996) have shown that—even in case of true association findings—a replication of the study will not necessarily have a high probability of yielding significance. If $p \approx 0.05$ in the initial study, for example, then the replication probability is close to only 50%! For a replication probability of 80%, one would need an association result that is significant at $p \approx 0.005$. While we do not recommend reporting results only for such low p values, these numbers certainly suggest that one should treat a study with a p value only slightly smaller than 0.05 as an interesting but still unconvincing support of the alternative hypothesis, and bear in mind that such a study has an almost equal chance of not being replicated in the future.

Cons: A strong word of caution is in order here. In the absence of any bias (such as PS), or if bias is present but corrected for, increasing the sample size reduces random error and increases power to detect any true disease-marker association. However if PS is present and not corrected for, increasing the sample size increases the confounding effects of PS and further elevates the rejection rate relative to the nominal significance level. In effect, when no true association exists, the test for association will have more power to detect PS! This explains why cryptic substructure can be a nuisance when large case–control studies are performed. In a nutshell, it is not only pointless but also detrimental to use very large sample sizes if the investigator does not correct for PS before carrying out a test for association.

Another consideration in epilepsy association studies is heterogeneity. We already know that some of the rare monogenic epilepsies can be caused by different genes, and there is no reason to believe that the more common epilepsies should not show heterogeneity as well. Heterogeneity increases random error in association studies, so that larger sample sizes are required to detect a given effect size. Alternative design techniques, however, may be applied to increase the effect size by choosing more homogenous subgroups of epilepsy. A variation that is involved in juvenile myoclonic epilepsy (JME), for example, might be detectable in a large sample of patients with idiopathic generalized epilepsy, but only a fraction of the sample size would have been necessary to detect this

effect if only JME patients were chosen. While this example may sound trivial, the point we want to make here is that more is not always better, and careful clinical evaluation might sometimes be a more useful strategy.

CHOOSING "BETTER" PHENOTYPES

In the last decade more than 50 association studies have been performed in epilepsy, but gene defects for common polygenic epilepsies have not yet been identified, leading to some skepticism about the association-study approach (reviewed by Tan et al., 2004). In addition to the methodological issues highlighted above, the selection of phenotypes that have been included in such studies deserves a comment; failures to replicate positive findings, or negative or false positive results, might also be related to phenotypic heterogeneity.

Idiopathic generalized epilepsies (IGEs), the most common target of association studies in epilepsy, include several subtypes based on age of seizures onset, predominant seizure types, and EEG features (ILAE 1989; Engel et al., 2001). Classification of individual patients is often challenging because existing categories are not mutually exclusive, and atypical traits or features common to more than one syndrome are frequent. For this reason a syndrome category of *IGEs with variable phenotypes* has been suggested as a new concept (Engel et al., 2001).

Classification of families for genetic studies is also difficult because multiple IGE syndromes often coexist within the same family, suggesting that IGEs share a common genetic background (Marini et al., 2004). It was therefore postulated that there might be genes that raise seizure susceptibility for IGE in general and that they act in combination with other genes that determine the specific seizure type (Durner et al., 2001). This concept of "seizure"—rather than "syndrome"—specific genes is also supported by Winawer et al. (2005) who found that families are more concordant for the same seizure type than IGE syndrome. These findings suggest that, for the purpose of genetic studies (and especially linkage studies), in addition to the syndrome approach, seizure types should also be taken into account when phenotyping patients or families.

While "lumping" different IGE phenotypes together might help to identify the more general genes, this approach will not be helpful with the identification of syndrome- or seizure-specific genes. But even with clinically homogeneous syndromes, causal heterogeneity in epilepsy is expected because of the effects of a number of "susceptibility" genes.

Unlike in linkage studies, we have no means to identify heterogeneity in a sample for association studies. In addition, heterogeneity will increase the sample size necessary to detect an association. We are therefore proposing to use clinical means to minimize heterogeneity and enrich a sample with patients as similar as possible. The goal should be not to collect "more" patients but "better" patients. This approach also means being more restrictive in the inclusion of patients—and patients with a syndrome diagnosis, but with atypical features, should be excluded. There are a variety of symptoms one could use to ensure more clinical similarity: For example, one could use EEG criteria like photosensitivity or polyspike wave to select for JME patients with or without photosensitivity, or for absence patients with strictly 3 Hz spike-wave and no polyspike wave. Alternatively, we could select only patients with a specific syndrome that show a certain pharmacospecificity, showing complete seizure remission under one specific drug. Another way to be more restrictive would be to choose a certain age-of-onset cutoff. Typical CAE cases may be limited to an age-of-onset between 5 and 7 years to minimize the chances that the patients have absence seizures that are in fact associated with Juvenile Absence Epilepsy on the upper age spectrum, or epilepsy with myoclonic-astatic seizures (Doose syndrome) at the lower end. This list is by no means exhaustive, and it will be a challenge for observant clinicians to increase the likelihood of success for future association studies and identification of epilepsy-related genetic variations by better defining the phenotype. Importantly, these criteria can then also be used for replication studies, thus ensuring that similar patients are studied and that nonreplication of positive results are not caused by inclusion of different forms of epilepsy.

In addition, we could be more restrictive not only in our choice of epilepsy cases but also controls. Instead of only requiring that the controls do not have epilepsy, we might collect controls that also have no family history of epilepsy. Or we might even subject them to an EEG examination to rule out the presence of epileptiform activity in the EEG.

Reducing genetic heterogeneity by careful, informed phenotype definition, including both clinical and EEG information is essential for gene identification because it allows division of the epilepsies into groups more likely to share susceptibility genes, and improves the power to find sound linkage or association. Unfortunately, IGEs might respond to a genetic model for which very large sample sizes, on the order of thousands, are required (Tan et al., 2004), with the consequent, unavoidable need for collaboration between centers and risk of underestimating heterogeneity and phenotyping inconsistencies.

Many researchers have voiced skepticism about an association approach to find genes involved in epilepsy, partially because of so many conflicting results. But before we throw out the baby with the bathwater, we believe we should first try to improve on this strategy by careful selection of cases and controls, avoiding possible bias, honest reporting of number of tests performed, and powering these studies enough so an effect can be detected. As we have shown, there will be many positive findings

that have a low chance of replication. However, that might not necessarily indicate that all of these findings are false positives. The challenge will be to identify which ones are true associations.

REFERENCES

Barratt C, Lai T, Nashef L, Valentin A, Fisniku L, Moran N, Asherson P, Makoff A. (2006) No association of single nucleotide polymorphisms in the μ-opoid receptor subunit gene with idiopathic generalized epilepsy 47:1728–1731.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B* 57:289–300.

Bird TD, Jarvik GP, Wood NW. (2001) Genetic association studies: genes in search of diseases. *Neurology* 57:1153–1154.

Cardon LR, Bell JI. (2001) Association study designs for complex diseases. *Nature Reviews Genetics* 2:91–99.

Colhoun HM, McKeigue PM, Davey Smith G. (2003) Problems of reporting genetic associations with complex outcomes. *Lancet* 361:865–872.

Devlin B, Roeder K. (1999) Genomic control for association studies. *Biometrics* 55:997–1004.

Devlin B, Roeder K, Wasserman L. (2001) Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology* 60:155–166.

Durner M, Keddache MA, Tomasini L, Shinnar S, Resor SR, Cohen J, Harden C, Moshe SL, Rosenbaum D, Kang H, Ballaban-Gil K, Hertz S, Labar DR, Luciano D, Wallace S, Yohai D, Klotz I, Dicker E, Greenberg DA. (2001) Genome scan of idiopathic generalized epilepsy: evidence for major susceptibility gene and modifying genes influencing the seizure type. *Annals of Neurology* 49:328–335.

Editorial. (1999) Freely associating. *Nature Genetics* 22:1–2.

Engel J Jr. (2001) International League Against Epilepsy (ILAE). A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ILAE Task Force on Classification and Terminology. *Epilepsia* 42:796–803.

Gorroochurn P, Hodge SE, Heiman GA, Greenberg DA. (2006a) A unified approach for quantifying, testing and correcting population stratification in association studies. (submitted).

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. (2006b) Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. *Genetic Epidemiology* 30:277–289.

Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. (1996) Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology* 33:175–183.

Halliburton R. (2003) *Introduction to population genetics*. Prentice Hall, Englewood Cliffs, New Jersey.

Healy DG. (2006) Case-control studies in the genomic era: a clinicians guide. *Lancet Neurology.* 5:701–707.

Hirschhorn JN, Altshuler D. (2002) Once and again-issues surrounding replication in genetic association studies. *The Journal of Clinical Endocrinology and Metabolism* 87:4438–4441.

ILAE. (1989) Commission on Classification and Terminology of the International League Against Epilepsy: Proposal for revised classification of epilepsies and epileptic syndromes. *Epilepsia* 30: 389–399.

Little J, Bradley L, Bray MS, Clyne M, Dorman J, Ellsworth DL, Hanson J, Khoury M, Lau J, O'Brien TR, Rothman N, Stroup D, Taioli E, Thomas D, Vainio H, Wacholder S, Weinberg C. (2002) Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *American Journal of Epidemiology* 156:300–310.

Lorenz S, Taylor KP, Gehrmann A, Becker T, Muhle H, Gresch M, Tauer U, Sander T, Stephani U. (2006) Association of BRD2 polymorphisms with photoparoxysmal response. *Neurosciences Letters* 400:135–139.

Marini C, Scheffer IE, Crossland KM, Grinton BE, Phillips FL, McMahon JM, Turner SJ, Dean JT, Kivity S, Mazarib A, Neufeld MY, Korczyn AD, Harkin LA, Dibbens LM, Wallace RH, Mulley JC, Berkovic SF. (2004) Genetic architecture of idiopathic generalized epilepsy: clinical genetic analysis of 55 multiplex families. *Epilepsia* 45:467–478.

Marks AR. (2003) Editorial policies and practices. *The Journal of Clinical Investigation* <http://www.jci.org/misc/jcipoli.pdf>.

Pritchard JK, Rosenberg NA. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 65:220–228.

Pritchard JK, Stephens M, Donnelly P. (2000a) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. (2000b) Association mapping in structured populations. *American Journal of Human Genetics* 67:170–181.

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

Sander T, Berlin W, Gscheidel N, Wendel B, Janz D, Hoehe MR. (2000) Genetic variation of the human μ-opoid receptor and susceptibility to idiopathic absence epilepsy. *Epilepsy Research* 39:57–61.

Spielman RS, McGinnis RE, Ewens WJ. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 52:506–516.

Tan NCK, Mulley JC, Berkovic SF. (2004) Genetic association studies in epilepsy: "The truth is out there." *Epilepsia* 45:1429–1442.

Winawer MR, Marini C, Grinton BE, Rabinowitz D, Berkovic SF, Scheffer IE, Ottman R. (2005) Familial clustering of seizure types within the idiopathic generalized epilepsies. *Neurology* 65:523–528.