A simple method for assessing the strength of evidence for association at the level of the whole gene

David Curtis[1]
Anna E Vine[1]
Jo Knight[2]

[1]Centre for Psychiatry, Queen Mary's School of Medicine and Dentistry, London E1 1BB, UK; [2]Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK

Introduction: It is expected that different markers may show different patterns of association with different pathogenic variants within a given gene. It would be helpful to combine the evidence implicating association at the level of the whole gene rather than just for individual markers or haplotypes. Doing this is complicated by the fact that different markers do not represent independent sources of information.

Method: We propose combining the p values from all single locus and/or multilocus analyses of different markers according to the formula of Fisher, $X = \Sigma(-2\ln(p_i))$, and then assessing the empirical significance of this statistic using permutation testing. We present an example application to 19 markers around the HTRA2 gene in a case-control study of Parkinson's disease.

Results: Applying our approach shows that, although some individual tests produce low p values, overall association at the level of the gene is not supported.

Discussion: Approaches such as this should be more widely used in assimilating the overall evidence supporting involvement of a gene in a particular disease. Information can be combined from biallelic and multiallelic markers and from single markers along with multimarker analyses. Single genes can be tested or results from groups of genes involved in the same pathway could be combined in order to test biologically relevant hypotheses. The approach has been implemented in a computer program called COMBASSOC which is made available for downloading.

Keywords: Fisher, significance, genetic marker

Introduction

A commonplace issue that arises when carrying out case-control studies to detect genetic association is that more than one marker within the same gene may support association. From a genetic point of view it may be expected that different markers may be in linkage disequilibrium (LD) with a pathogenic variant and from a biological point of view it may be expected that different variants within the same gene may have a role in influencing risk of disease. Often the hypothesis of interest is whether variants in a given gene influence risk rather than whether one particular marker demonstrates association. Hence it would be desirable to combine information from multiple markers in order to obtain an overall measure of the evidence implicating a gene. As has been discussed (Neale and Sham 2004), this issue is perhaps especially pertinent in the context of GWA studies. If the situation arises where a number of markers within a single gene achieve modest levels of significance then most people would agree that this finding would be of more interest than if the same number of markers achieved the same results but were randomly positioned with respect to each other.

Typically, an association study claiming to find evidence to support the involvement of a gene will present results obtained from several or many markers in the vicinity.

Correspondence: David Curtis
Adult Psychiatry, Royal London Hospital, Whitechapel, London E1 1BB, UK
Tel +44 20 7377 7729
Fax +44 20 7377 7316
Email david.curtis@qmul.ac.uk.

A few single markers may individually produce small *p* values and results from some multimarker methods, using logistic regression or inferred haplotypes, may be presented as offering additional support. Varied numbers of markers in different combinations will have been studied and results from the analyses yielding the most positive results will be presented. There may be an attempt to deal with multiple testing issues by carrying out simulations in order to obtain the empirical significance of the most highly significant result. However we argue here that the main point of interest is not the true statistical significance of only the most strongly positive analysis but rather the inference to be derived from the overall combination of results obtained from different markers and methods. It is this combination of results, in the form of *p* values from different single marker and multimarker tests, which is usually presented by the authors with the tacit invitation that readers use their own judgement and intuition to decide on the strength of the evidence implicating the gene in question. It would be helpful to have a formal method to support this process.

A number of complexities need to be dealt with. Firstly, markers within a gene do not represent independent sources of information since some will be in LD with each other. Also, there may be different variants influencing risk, perhaps to different extents. If this is so then alleles of some markers may show association through their proximity to one variant while other markers may detect the effect of another variant. Alternatively, different haplotypes of the same marker set may be associated with different variants. Some markers within the same gene may demonstrate little or no LD with each other and hence be relatively independent. Markers some distance from the coding region may nevertheless detect association. There may be a relatively large number of markers to deal with and methods which involves combining all into a conventional multi-marker analysis (Chapman et al 2003, 2007; Clayton et al 2004) may be impractical, because of the large number of parameters involved, and/or inappropriate, because different variants may produce different patterns of association with different subsets of markers.

One early approach to tackling this issue was to consider combining results from groups of neighboring markers which were close enough to each other to be in LD (Zaykin et al 2002). This resulted in a series of *p* values produced from overlapping marker sets forming a sliding window analysis but did not produce an overall statistic at the level of the whole gene. A subsequent development (Chen et al 2006) considered combining results from analysis of single SNPs with one overall haplotype analysis. Other approaches combined results from either single marker analyses (Hoh et al 2001; Potter 2006) or results from different multimarker analyses using sliding windows incorporating a weighting scheme for markers flanking the central marker of each window (Yang et al 2006). The evaluation described in the first of these studies (Potter 2006) showed that combining *p* values according to the method of Fisher (Fisher 1925) produced good power compared with other approaches. A method has been proposed to use extreme-value distributions to evaluate the significance of results over blocks of markers (Dudbridge and Koeleman 2004) but it is not clear that this could readily be applied to the variety of different methods which are used to evaluate the evidence implicating a particular candidate gene. Here we present a natural development of these ideas which allows the assessment of a whole gene. It differs from previous methods in that it uses information from both multiple markers and multiple methods of analysis. Information can be combined from single marker analyses along with multimarker analyses using different numbers of markers, which may be biallelic or multallelic, and based on haplotypes or locus-scoring methods. No matter how many different methods are applied, one can still arrive at an overall *p* value which provides a measure of the strength of evidence supporting the hypothesesis that one or more variants in the gene influence susceptibility to the phenotype being studied.

Method

The approach consists of two stages. The first is to combine the evidence for association and the second is to assess the strength of the evidence.

The method we use for combining *p* values is that due to Fisher (1925). This is based on the observation that, if *n* independent tests are made of the same hypothesis, then $X = \Sigma(-2\ln(p_i))$ is distributed as a χ^2 with $2n$ df. The *p* values to be combined could be obtained from a set of single marker analyses or could come from both single marker and multimarker analyses. The summative measure obtained, X, could be taken to provide a combined measure of the strength of evidence in favor of association for a group of markers except that we do not expect the contributions to be independent.

This is dealt with by the second stage of our procedure which is simply to use permutation testing to assess the empirical significance of X. If we keep the multimarker genotypes intact and permute case-control labels then this will fully deal with all the interdependencies of the markers due to LD between them and of interdependencies between

the methods of analysis. Among other things it may also mitigate the effects of over-correcting for multiple markers (as could occur if a Bonferroni correction were applied), some of which may be scarcely informative. The procedure we propose for obtaining an empirical significance is sequential Monte Carlo testing (Besag and Clifford 1991). When carrying out permutation testing, rather than setting the number of permuted replicates, n, to a fixed number one instead sets a target for r, the number of times that a permuted replicate should exceed the test statistic obtained from the real dataset. Typically the target for r might be set to a value of 10 or 20. One would also set some maximum value of n to ensure that the procedure did eventually finish. If the target value for r is reached then the empirical significance is given by $p = r/n$ while if the target is not reached before n reaches its maximum value the empirical significance is given by $p = (r + 1)/(n + 1)$, as used in conventional Monte Carlo testing (North et al 2003a). The sequential approach produces a very valuable increase in speed of permutation testing when the p value to be estimated turns out to be non-significant. If there is no association present then the number of permutations expected to be performed before the target is reached is approximated by $r + r \log((n + 1/2)/(r + 1/2))$ (Besag and Clifford 1991). For example, with a target of $r = 10$ and $n = 9999$ then one may expect to perform 39.8 permutations, achieving a 250-fold speed increase compared with using the conventional method. By permuting the multimarker genotypes against phenotype this approach can be trusted to yield the correct Type 1 error rate when the null hypothesis is true.

To summarize, we propose that to obtain an overall measure for the strength of evidence supporting involvement of a gene which has been typed with a number of markers subjected to different single locus and multilocus methods of analysis one first derives $X = \Sigma(-2\ln(p_i))$ and then assesses the empirical significance of X using permutation testing.

In order to provide a demonstration of the approach in practice, we applied it to a publicly available case-control dataset. This consisted of consisted of 270 subjects with Parkinson's disease and 271 controls genotyped for a GWA study using the Illumina Infinium I and Infinium II assays (Fung et al 2006) These genotypings were downloaded from the Coriell Institute (http://ccr.coriell.org). There has been a previous report that two different mutations within the HTRA2 gene may be associated with Parkinson's disease (Strauss et al 2005). According to the UCSC browser (http://genome.ucsc.edu/cgi-bin/hgGateway), HTRA2 is located on chromosome 2 at 74610040-74614191. We selected 19 SNPs spanning this region ranging from rs6718621 at 74512208 to rs10170219 at 74715172 and calculated individual p values testing for association with each marker using the SCANASSOC program (Curtis et al 2006). In addition to single marker analyses we carried out haplotype-based tests for association using consecutive sets of two or three markers. We then applied the new approach to assess the overall evidence for association obtained from this group of 19 markers. We set a target of 10 for r, the number of permuted datasets to achieve the value obtained from the real one, and we set a maximum number of permutations, n, to be 9999.

Results

The results from the tests of the individual markers are shown in Table 1. It can be seen that one marker, rs2241027, is significant at $p = 0.04$ and that two others yield p values below 0.1. One three-marker analysis has a test-wise significance of 0.03. We combined all 54 values according to the formula $X = \Sigma(-2\ln(p_i))$ and obtained a value of 149.4. Taking this as a χ^2 statistic with 108 degrees of freedom would produce a nominal p value of 0.005. However, when we carried out permutation testing the target number of 10 permuted datasets to produce this value or higher was reached after only 62 permutations, corresponding to an empirical significance of $10/62 = 0.16$.

Discussion

The approach we propose seems simple and to have face validity. It adequately deals with the issues of non-independence between markers and methods of analysis while allowing the combining of information from many markers from different regions of the same gene. There may be some benefit in considering it in relation to other approaches for combining evidence from diverse sources. The philosophy underlying the Bonferroni correction and related procedures such as the estimation of the false positive report probability (Wacholder et al 2004) is that one is carrying out a number of unrelated experiments and one wishes to test whether for at least one of them the alternative hypothesis may be true. The philosophy of Fisher's approach is that one is carrying out multiple independent experiments to test a single hypothesis. Notionally, one may then expect that the same effect will be present in all experiments although stochastic factors will impact on the results one obtains in practice. Thus one may expect that some studies may yield significant results while others may, through chance or small sample size, be formally non-significant.

Table 1 Markers spanning HTRA2 showing individual p values obtained for tests for association with Parkinson's disease

Marker	Position	P value		
		Single marker	Two markers	Three markers
rs6718621	74512208	0.104	0.223	0.343
rs6751601	74536436	0.224	0.084	0.501
rs2240444	74553279	0.061	0.225	0.357
rs2268424	74560023	0.133	0.303	0.596
rs2268420	74566516	0.689	0.564	0.728
rs2268418	74576122	0.166	0.505	0.623
rs7556852	74581170	0.200	0.392	0.691
rs6746854	74593157	0.218	0.326	0.326
rs1063588	74602033	0.149	0.149	0.259
rs1047911	74611433	0.149	0.259	0.136
rs6707475	74622146	0.206	0.109	0.141
rs2301984	74632710	0.155	0.218	0.404
rs2240442	74645428	0.097	0.255	0.260
rs3806607	74647269	0.212	0.152	0.034
rs2241027	74670321	0.041	0.175	0.369
rs6707302	74673077	0.273	0.637	0.617
rs7562200	74688601	0.841	0.326	0.515
rs11126435	74701355	0.126	0.465	
rs10170219	74715172	1.000		

Nevertheless one will tend to see that the *p* values obtained over all experiments are smaller than would be expected by chance. When interpreting data from markers around a single gene one faces a hybrid situation. One expects that some markers may provide information regarding the main hypothesis, that the gene concerned affects the phenotype studied, while other markers will not be in LD with functional variants and hence will behave as unrelated sources of essentially random effects. One way to model this situation would be to carry out logistic regression analysis with each marker being treated as an independent variable contributing to risk (Chapman et al 2003), although it is not clear the extent to which significance testing based on asymptotic distributions would be appropriate if more than a few markers were included in such an analysis. Certainly, what we notice in practice is that authors report the best results they have obtained from single marker and multimarker analyses, generally without any formal attempt to consolidate the overall evidence implicating a gene. Our method of combining all results from all sources and carrying out permutation testing does provide a means to obtain such a summary *p* value.

Although we are unable to find any published account, it appears that a somewhat similar method to ours may be implemented in Shaun Purcell's PLINK program (Purcell et al 2007) as described in the on-line documentation (http://pngu.mgh.harvard.edu/~purcell/plink/anal.shtml#set). According to the documentation accompanying version 0.99q (3 March 2007), this can carry out analyses on subsets of markers selected from a set designated by the user. The subset size is varied between a minimum and maximum size also specified by the user and the best result obtained is defined in terms of the sum of the largest chi-squared statistics from a subset of each size. The overall significance is then evaluated using a permutation procedure. We suspect that our method would be similar to setting both the minimum and maximum subset size to be equal to the size of the whole set. That is, one would simply sum the chi-squared results for all markers. However the documentation implies that one should avoid doing this by setting the maximum subset size to a "reasonable number" in order to avoid performing an "unnecessary number of tests". If the minimum and maximum sizes differ then in fact additional tests are performed for the different sizes. Our approach explicitly addresses the possibility that different kinds of analysis might be used. The software we have implemented allows incorporation of locus-based logistic regression as well as haplotype-based analyses. In principle other methods of analysis, for example neural network analysis (North et al 2003b) or haplotype clustering methods (Knight et al 2008), could be accommodated. Our approach defines in advance the markers of interest and

takes a summary statistic derived from all their p values simultaneously. Likewise, previous work suggests that, at least in some situations, more powerful tests will result if only p values below a certain threshold are combined (Zaykin et al 2002). Once again, the choice of threshold is arbitrary and it is not clear that using this truncation will always be of benefit. The exploration of the advantages and disadvantages of each approach could be the subject of further investigation.

Our example application does not provide support for association between Parkinson's disease and HTRA2. Different conclusions might have been drawn had there been a stronger prior hypothesis, for example if the three markers with $p < 0.1$ had been specifically implicated in other studies. At the level of the gene, however, our overall result is negative. There appear to be more small p values than would be expected by chance (as is clear from the $\Sigma(-2\ln(p_i))$) so a naïve interpretation might have been that these markers did support association. However once we apply our permutation we can see that, because of the non-independence of the p values, in fact the results are well within chance expectation. This demonstrates the value of our approach in being able to summarize the available evidence.

A number of extensions to this basic approach could be developed. We should begin by pointing out that even as it stands the method can combine information from biallelic and multiallelic markers. It can also combine information from both single marker analyses and multimarker analyses. Thus one might wish to treat some sets of markers as being suitable for haplotype analysis and combine the information from these with results from other markers or groups of markers. As demonstrated in the example above, one can also combine results from single marker and multimarker analysis of the same markers. That is, if one had 4 markers one could combine the 4 single marker p values along with a p value obtained from haplotype analysis of all of them. Using multiple methods of analysis may risk reducing power somewhat but the overall significance level obtained remains valid.

Other ways could be considered to combine the individual p values. For example, more weight could be given to markers within coding regions or those closer to rather than further from the gene or those having been implicated in previous studies. Again, the permutation testing will ensure that whatever method is used to combine them the empirical significance level will still be valid. Results from functionally related groups of genes could be combined. This would provide evidence to implicate a particular pathway or system rather than an individual gene.

We should note some situations in which the empirical significance level would not be valid. The main principle is that the p values to be combined must not be selected on a post hoc basis. For example, one must not notice that a particular intron contains a number of interesting results and then combine the results just from that intron. One cannot elect to include markers from some distance away after seeing that some appear to support association. One cannot perform a number of different multimarker analyses and then include the results from only the most significant ones. One can apply this approach to a gene which appears interesting based on the fact that a number of markers within it appear to show some evidence for association but only if one then proceeds to make a standard multiple-testing correction for all the other genes for which genotypes were obtained.

We acknowledge that although our approach may appear theoretically attractive we are not currently able to present clear evidence regarding its power compared with other methods. This is because it is intended to deal with a situation which is biologically plausible – that different mutations in the same gene might each have an effect on a given phenotype – but for which real data are lacking and for which plausible computer simulations would be technically difficult. One would need to model datasets in which multiple mutations occurred within the same gene along with the complex and realistic LD relationships for markers around each mutation. We have previously studied such models in the context of a single mutation (North et al 2006) but have not as yet produced a procedure to carry out systematic studies of the simulated effects of multiple mutations. We do not expect that the approach would be any more powerful than pre-existing methods in the simple situation of a single mutation. Although we cannot claim to have demonstrated that the approach is necessarily more powerful than other methods, we are confident at least that the permutation procedure means that the overall result is valid, that is that the Type 1 error rate is correct. This means that our approach does at least provide a way of summarizing the available evidence implicating a particular region rather than having to rely upon the reader's subjective judgment based on a number of non-independent p values obtained from different analyses.

The method for combining results from different analyses has been implemented in the COMBASSOC program, which is available along with the other programs to support GENECOUNTING (Zhao et al 2002), available from our website at: www.mds.qmul.ac.uk/statgen. Analyses can consist of any number of single marker tests, multimarker

haplotype analyses and multimarker locus-wise analyses using logistic regression. If desired, different subsets of markers can be selected for different analyses. No matter how many different tests are performed, results from all are combined to produce one overall measure of the strength of evidence in favor of association and the empirical significance of this is derived using permutation testing, providing a single overall *p* value.

We hope that the approach outlined will prove attractive and practical. It provides a simple and intuitive way to provide some objective assessment of the overall evidence for association produced by a group of markers. We consider such an approach to be preferred to the widespread practice of quoting the individual significance of a number of different single marker and/or multimarker analyses and leaving it to the reader to form some kind of judgement as to the implication of the results.

Acknowledgments

AEV was supported by Wellcome Trust Project Grant, Grant No. 076392. JK was supported by an MRC Bioinformatics Training Fellowship, Grant No. G0501329.

Disclosures

The authors disclose no conflicts of interest.

References

Besag J, Clifford P. 1991. Sequential Monte Carlo p-values. *Biometrika*, 78:301–4.

Chapman J, Clayton D. 2007. One degree of freedom for dominance in indirect association studies. *Genet Epidemiol*, 31:261–71.

Chapman JM, Cooper JD, Todd JA, et al. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56:18–31.

Chen BE, Sakoda LC, Hsing AW, et al. 2006. Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genet Epidemiol*, 30:495–507.

Clayton D, Chapman J, Cooper J. 2004. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol*, 27:415–28.

Curtis D, Knight J, Sham PC. 2006. Program report: GENECOUNTING support programs. *Ann Hum Genet*, 70:277–9.

Dudbridge F, Koeleman BP. 2004. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am J Hum Genet*, 75:424–35.

Fisher RA. 1925. Statistical methods for research workers, 13 ed. London: Oliver and Boyd.

Fung HC, Scholz S, Matarin M, et al. 2006. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*, 5:911–6.

Hoh J, Wille A, Ott J. 2001. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res*, 11:2115–9.

Knight J, Curtis D, Sham PC. 2008. CLUMPHAP: a simple tool for performing haplotype-based association analysis. *Genet Epidemiol*, 32:539–45.

Neale BM, Sham PC. 2004. The future of association studies: gene-based analysis and replication. *Am J Hum Genet*, 75:353–62.

North B, Sham PC, Knight J, et al. 2006. Investigation of the ability of haplotype association and logistic regression to identify associated susceptibility loci [online]. *Ann Hum Genet*, 70:893–906.

North BV, Curtis D, Sham PC. 2003a. A note on calculation of empirical P values from Monte Carlo procedure. *Am J Hum Genet*, 72:498–9.

North BV, Curtis D, Cassell PG, et al. 2003b. Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Ann Hum Genet*, 67:348–56.

Potter DM. 2006. Omnibus permutation tests of the association of an ensemble of genetic markers with disease in case-control studies. *Genet Epidemiol*, 30:438–46.

Purcell S, Neale B, Todd-Brown K, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81:559–75.

Strauss KM, Martins LM, Plun-Favreau H, et al. 2005. Loss of function mutations in the gene encoding Omi/HtrA2 in Parkinson's disease. *Hum Mol Genet*,14:2099–11.

Wacholder S, Chanock S, Garcia-Closas M, et al. 2004. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96:434–42.

Yang HC, Lin CY, Fann CS. 2006. A sliding-window weighted linkage disequilibrium test. *Genet Epidemiol*, 30:531–45.

Zaykin DV, Zhivotovsky LA, Westfall PH, et al. 2002. Truncated product method for combining P-values. *Genet Epidemiol*, 2002, 22:170–85.

Zhao JH, Lissarrague S, Essioux L, et al. 2002. GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics*, 18:1694–5.