European Journal of Human Genetics (2006) 14, 529–534
© 2006 Nature Publishing Group All rights reserved 1018-4813/06 $30.00
www.nature.com/ejhg

ARTICLE

Use of phenotypic covariates in association analysis by sequential addition of cases

Stuart Macgregor*,[1], Nick Craddock[2] and Peter A Holmans[1]

[1]*Biostatistics and Bioinformatics Unit, Cardiff University, Cardiff, UK;* [2]*Department of Psychological Medicine, Cardiff University, Cardiff, UK*

Optimal use of phenotype information is important in complex disease gene mapping. We describe a method, sequential addition, for the analysis of case–control data by taking into account of a quantitative trait that is measured in cases but not in controls. The method also provides an estimate of the best phenotype definition for future studies. We demonstrate proof of principle, using an example of incorporation of age-at-onset data into a study of a small sample for association between APOE and late-onset Alzheimer's disease. The sequential addition method finds evidence of association when conventional case–control methods fail. We also illustrate the use of the method for taking account of a dimensional measure of psychosis in a study of the schizophrenia susceptibility gene, dysbindin, in bipolar disorder.

European Journal of Human Genetics (2006) **14,** 529–534. doi:10.1038/sj.ejhg.5201604; published online 15 March 2006

Keywords: phenotype definition; case only; quantitative trait; bipolar disorder; late-onset Alzheimer's disease; ordered subset analysis

Introduction

Association studies using either quantitative or qualitative traits are widely used for molecular genetic studies of complex disorders. In some circumstances, the available data will comprise a set of cases and controls where the cases (but not the controls) are measured for a quantitative trait. This situation typically arises when the quantitative phenotype is undefinable in controls; for example, when the trait is age at onset (AAO) of the disease under study. This case-only quantitative trait may be critical in defining a more homogeneous subset of the available cases. It is now well established that appropriate phenotype definition is crucial for gene mapping in complex human disease.[1–5] The problem of appropriate phenotype definition is particularly acute in psychiatric disease with many clinical variables being only defined in case individuals but not in controls. For example, in Bipolar disorder, scores on the Bipolar Affective Disorder Dimension Scale (BADDS[6]) may be useful in more accurately defining the phenotype.

Since the controls have no quantitative trait measure, conventional approaches for quantitative association analysis (such as regression) cannot be used. *Ad hoc* procedures such as selecting cases with trait values below or above a particular threshold can be applied, but it is often unclear *a priori* what threshold to use. We propose a method, sequential addition (SA), for analysing case–control data when there is a quantitative trait measured only in cases and where no threshold is known. A simulation-based procedure is used to maintain an appropriate false positive rate. The SA method is related to the ordered subset analysis (OSA) described by Hauser *et al.*[7] The OSA method allows incorporation of quantitiative covariate information to be used to select informative subsets of families for linkage analysis. The SA procedure we describe here

*Correspondence: Dr S Macgregor, Current address: Queensland Institute of Medical Research, Brisbane, Australia.
Tel: +61 7 3845 3563; Fax: +61 7 3362 0101;
E-mail: stuart.macgregor@qimr.edu.au

Received 7 November 2005; revised 24 January 2006; accepted 8 February 2006; published online 15 March 2006

performs the same role for case–control samples with case-only covariate information.

Here, we demonstrate the potential utility of the SA method with a previously studied Alzheimer's disease (AD) data set and AAO data. The method is then applied to a Bipolar disorder data set and a quantitative measure of psychosis.

Materials and methods

The principle of the SA method is to maximise the significance of the association test between a set of cases and controls by sequentially adding case individuals in ascending or descending order according to the value they have at the quantitative trait. The association test is repeated for data sets with increasingly large numbers of cases included. If the quantitative trait is important in defining the phenotype–genotype relationship at the locus in question, then one end of the distribution of trait values will contribute disproportionately to the association signal. Since many tests are conducted in the procedure, we appeal to computer simulation methods to determine the overall significance of the finding. If the effect size in a subset of the individuals is larger than the effect size in the whole sample then this may offset the disadvantage of multiple testing.

The procedure (for a descending SA analysis) is

1. Sort cases by the quantitative trait.
2. Add the case with the (next) highest trait value to the sample for analysis.
3. Calculate the relevant test statistic for cases *versus* controls for marker(s)/haplotype(s) of interest.
4. Repeat steps 2 and 3, adding in cases sequentially and recalculating.
5. Store the smallest nominal *P*-value from all the tests done.

Since comparing a very small number of cases with the control set is unlikely to yield an interesting result, we propose adding 10 cases (with highest trait value) in the first iteration of step 2. For some quantitative traits (eg AAO), the lower end of the trait distribution may be thought to be most important in defining a genetically homogeneous subset. In these cases, it may make sense to sequentially add cases lowest trait value to highest trait value. To establish the significance of the *P*-value obtained from the above procedure, the set of genotypes are permuted among the whole sample (cases and controls) and the procedure is repeated a large number of times. The empirical *P*-value, correcting for the multiple tests done, is the proportion of permutation replicates that yield *P*-values smaller than that observed in the actual data set.

The analysis method in step 3 will vary depending on the test of interest. One basic test is an allelic test of association at a single SNP, performed using a χ^2 test (with empirical *P*-values when cell counts are small) on a contingency table. Effect size estimates can be calculated based on the odds ratio (OR) from the contingency table. Confidence intervals (CI) on the ORs can be obtained using a standard formula.[8] Haplotype-based SA can be implemented by utilising a haplotype based test in step 3.

The quantitative trait value at which the nominal *P*-value minimises (best cutoff) will be of interest because it will allow future studies to focus on a particular phenotype definition and thereby maximise power. To calculate a CI on the estimate of the best trait value cutoff a bootstrap procedure can be applied. Bootstrapping is where new samples are generated by randomly sampling case individuals with replacement.[9] The best cutoff is then recalculated on the new sample. Repeating this procedure over a large number of replicates and observing the 2.5 and 97.5 percentiles of the cutoffs allows the construction of an empirical 95% CI. We illustrate the method with the data sets described below using *R*.[10] In each of the cases below 500 bootstrap replicates were generated.

The SA method was applied to a small subset of a late-onset AD data set[11] for the covariate AAO, clearly measurable only in cases. The importance of AAO in the definition of the AD phenotype is well established[12,13] and we demonstrate here the potential utility of the SA method as a proof of principle. Individuals were added in increasing order of AAO. There are 40 case individuals in our sample with AAO values ranging from 65 years to 98 years. Forty control individuals were available. All individuals were typed for the established AD risk locus APOE with the alleles coded as the two allele system epsilon4/not epsilon4.

We then examined a set of data used in a case–control association analysis of the dysbindin gene in bipolar disorder.[14] The available case data were 592 individuals with scores for the BADDS[6] psychosis dimension (mean 34.1, standard deviation 29.2). Psychosis is measured on a 0–100 scale and has been shown to have a significant familial component.[15,16] A total of 1251 control individuals were available. All individuals were typed for rs2619538 from Raybould *et al*.[14] Dysbindin has previously been implicated in association studies of schizophrenia;[17] SNP rs2619538 demonstrated association in our schizophrenia sample, which was recruited from the same clinical and geographical population and using the same methodology as our bipolar disorder sample.[18] In the light of this positive result in schizophrenia, we hypothesised that the BADDS psychosis dimension may be important in refining the phenotype in our bipolar sample. Since it is unknown *a priori* what definition of psychosis is likely to be most useful we apply the SA procedure over the full range of psychosis values.

An *R* script that implements the method described here is available on request from the corresponding author.

Results

Applying the standard case–control association test to the AD data yielded a *P*-value of 0.11 for the effect of the APOE locus on AD. The SA procedure with AAO yielded a permutation corrected (100 000 replicates) empirical *P*-value of 0.0082. The minimum *P*-value ($P = 0.0009$, not corrected for multiple testing) was obtained when only individuals with AAO<73 years and 4 months were included in the analysis (12 individuals met this criteria). Older individuals do not improve the significance of the result because, although including them increases the sample size, the frequency of the epsilon4 allele in this older subset is close to the frequency in the control group. The 95% bootstrap CI for the best cutoff was (70.8–79.3 years).

A graph showing the estimated OR for the increasingly large subsets of AD patients is shown in Figure 1. At the left-hand side of the graph, 10 cases are included in the OR calculation while on the far right-hand side the full set of cases is included. The error bars show the 95% CI on the OR; for clarity the upper error bar has been omitted. Although the CI are wider (sample size is smaller) on the left-hand side of the graph, due to the increased OR, the bottom of the 95% CI is above 1.25 for AAO values in the range (71.0, 76.8).

Applying the standard case–control association test to the full 592 individual bipolar data set yielded a *P*-value of 0.34 for the effect of the rs2619538 SNP on bipolar disorder; this was consistent with the non-significant result reported by Raybould *et al.*[14] The SA procedure yielded a permutation corrected (10 000 permutations) *P*-value of 0.020 when the cases were added highest to lowest (most severe psychosis first). The minimum *P*-value ($p = 0.0022$) was obtained when only individuals with dimension score ⩾45 were included in the analysis (195 individuals met this criteria). The 95% CI for the best cutoff was (23–80), with 50% of bootstrap replicates falling in the range (43–45).

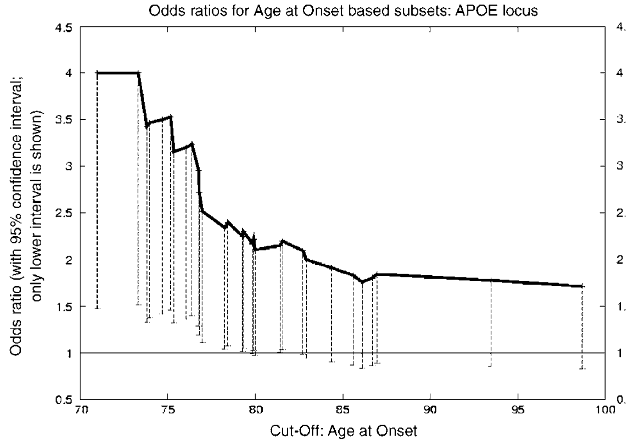

Figure 1 SA with AD and AAO.

Discussion

In summary, we have described a method, SA, for the analysis of case–control data where there is quantitative trait information available in cases only. SA is a replacement for *ad hoc* procedures such as arbitrarily choosing a cutoff for the definition of caseness. The method employs permutation testing to account for the multiple testing and also allows the estimation of the best phenotype definition (trait cutoff) for future studies.

We showed that, in our small proof-of-principle AD sample, the SA method could demonstrate a significant association by taking account of the covariate when the simple analysis was nonsignificant. The best trade-off between sample size and effect size at the APOE locus occurred when individuals aged 73.3 years (95% CI 70.8–79.3) were included in the analysis. In addition to this sample of 40 AD cases and controls, we had another seven independent sets of 40 cases/controls available (these were not used in the main analysis because such large samples give very small *P*-values, irrespective of the analysis method applied). We repeated the SA procedure in each of these data sets; the minimum *P*-value occurred at ages between 73.9 and 79.0. All of these fall within the CI calculated on the initial set of 40 cases/controls. We stress that, particularly in small samples, the range of trait values (in this case age) may not necessarily be truly representative of the trait values measured in another sample. For example, in the case of AD, some data sets only include individuals with particularly early onset and we would not expect the best trait cutoff in such studies to necessarily be the same as the trait cutoff found to be optimal here. With all eight AD data sets pooled (320 cases and 320 controls), the best trait cutoff was 76.8 (95% CI 72.7–79.3). We note here that the AD example was given primarily as a proof-of-principle and that we chose to discard information on the age measures for the control individuals. There are survival analysis techniques that explicitly deal with the situation where some of the individuals (ie the controls here) have not reached the age of onset or are censored. Further details of such techniques are given elsewhere.[19]

In the bipolar disorder study of dysbindin, the effect size increased considerably in the subset defined by psychosis. The psychosis-based SA procedure yielded a significant *P*-value (0.02) for the overall association of dysbindin with the SNP rs2619538. This result is particularly interesting given previous results for dysbindin in schizophrenia; further discussion of the role of psychosis in psychiatric disease definition is given elsewhere.[5]

The success of the SA procedure in other samples will depend on the relationship between sample size and effect size. The effect size in a subset of the cases must increase sufficiently compared with the effect size seen in the whole sample. A simple example will demonstrate that the effect size in the subset need not be substantially larger than that seen in the full case set. For example, suppose there are 460 cases and controls with a disease allele frequency of 0.4. Assuming the causative locus is typed, an OR of 1.3 (multiplicative model) is required for 80% power at the 5% level. Similar power would be achieved if the number of cases was half that of the original sample, but the effect size in the new subset was 1.52. Note to calculate this approximate comparison, we assume that the equivalent of nine independent tests were carried out to find the best subset and that hence the 80% power is obtained at the 0.555% level (Bonferroni correcting for nine tests) instead of the 5% level. To derive this approximate value of 9 for the 'equivalent number of independent tests', we assume a data set similar to the psychosis data described in the results section above. This value of 9 follows because the permutation derived *P*-value (which corrects for the multiple tests done) is approximately nine times larger (0.020/ 0.0022$\sim$9) than the minimum asypmptotic *P*-value (which does not correct for multiple testing). The actual number of tests performed for the psychosis data was 60, but because of the overlap in the tested subsets there was substantial correlation between tests. With other data sets the multiple testing penalty will vary depending on the covariate of interest. Clearly, there are a wide range of possible models and power will vary depending on the (unknown) genetic model. The selected subset of cases will vary depending on the quantitative covariate available and the power of the SA approach will depend upon the utility of this covariate in identifying a suitably homogeneous subset. When there is substantial genetic heterogeneity, we may expect a subset of the cases to be affected as a result of their genotypes at loci distinct from the locus under study; these individuals will be more likely to carry the 'wild-type' allele than the 'disease' allele at the locus of interest. Useful covariates, therefore, will be ones that identify these subsets and hence allow efficient removal of these uninformative individuals.

An alternative to the SA procedure is to perform the quantitative trait association analysis in cases only. This can be applied through the use of programs such as qtphase[20] in the case of haplotypes or through the use of standard regression procedures in the case of single SNPs. It is worth noting that this regression-based approach is a test of whether the association depends upon the trait not of overall association. The SA procedure can be modified to perform an analogous test of whether the association depends on the trait by changing the way in which the permutation procedure is implemented (ie by permuting among cases only). However, since the main interest is commonly an initial test of overall association, here we have implemented the joint test of association and quantitative effect (ie permuting the genotypes among the whole sample). A positive result is hence indicative of there being a significant association in a subset of cases defined by the covariate. In itself, this does not indicate that the association depends upon the covariate within the resultant subset. If one is interested in a specific test of whether the association depends upon the trait, then the relative merits of an SA-based procedure and a procedure based on linear regression in cases only will depend upon how the allele frequency in cases changes with increasing values of the quantitative trait. If there is a roughly linear relationship between the quantitative trait and the allele frequency in cases then the regression procedure may be apt. If there is a relatively sharp cutoff point in the quantitative trait where the allele frequency changes dramatically then the SA procedure would be expected to perform substantially better than the regression procedure. It is possible to use more flexible regression techniques, such as fractional polynomials. However, the optimal choice of model will often be unclear.

In the applications of the SA procedure given here, there was a prior hypothesis of which end of the quantitative trait distribution was most relevant to the locus of interest. For AD, early onset cases were of primary interest and in bipolar disorder, individuals with high psychosis values were thought to be most important in defining the phenotype. However, for other quantitative traits, there may not be a clear prior hypothesis. In such cases, we would recommend that individuals are sequentially added highest first and then lowest first. This means that, if the optimal subset of cases includes mainly individuals with only high trait values or only low trait values, these subsets will be tested. This modification can then be repeated in the permutations, hence ensuring appropriate correction for the multiple tests carried out. If more than one trait is used to help define the phenotype, then an appropriate correction for multiple testing will also be necessary.

The SA method can be applied to multiple markers in a number of ways. Firstly, if a few markers are of interest, a single global haplotype based test such as that implemented in cocaphase[20] can be applied. Alternatively, the SA approach can be applied to each locus individually. In this case, the test statistic is calculated for each marker for each subset. The permutation procedure described in the Methods section is then applied to obtain the significance of the highest test statistic from any of the markers considered. Modifications of this procedure in which a series of sliding window haplotype tests are applied could also be utilised.

The SA approach is flexible in that a number of possible tests of association can be conducted. Possibilities include allele, genotype and haplotype based tests. The assumptions made in each test should be carefully considered. In the case of allelic tests, random mating is assumed and tests for deviation from Hardy–Weinberg equilibrium (HWE) should be conducted. Such HWE tests would typically be conducted on the full set of cases and controls (as a screen for genotyping errors). In addition to such tests, we would recommend that a separate HWE check is also conducted in the subset of cases found to be most significant for a given marker. In the case of haplotypic tests, where haplotype frequencies must be estimated, investigators should be aware that the estimated haplotypes may vary across the different possible subset of the case data. To ensure that the uncertainty in haplotype frequency estimation is appropriately taken into account in the test for association, likelihood based tests, such as that implemented in programs such as cocaphase[20] should be used. The effect of estimating the haplotype frequency from varying numbers of individuals will be minimized because likelihood ratio based tests include an estimate of haplotype frequencies derived from the full set of cases and controls (in addition to the frequencies in cases and controls separately). In contrast, haplotype-based tests that simply compare the estimated haplotype frequencies in cases *versus* controls are inappropriate.

A related method for sequentially adding in families into linkage analysis (ordered subset analysis or OSA) was described by Hauser *et al.*[7] OSA subsets the available family data according to the values for a particular covariate. Although this leads to analysis results based on only a subset of the data, the results in Hauser *et al* and in a number of subsequent publications (eg BHF Family Heart Study Research Group[21]) demonstrate that in many cases, selecting a genetically homogeneous subset leads to improved results. The OSA approach shares desirable properties with the SA approach we describe here. Both approaches require no *a priori* specification of the cutoff required to select a homogeneous subset of the data and both provide guidance for the selection on individuals/families for confirmatory studies. We note also here that the SA procedure described above can be simply modified for application with trios measured for a suitable quantitative trait.

More generally, the characteristics of maximally selected statistics have been examined in the literature. A commonly examined case is where the classification of individuals as either case or control is determined by an underlying quantitative trait. The total number of individuals in such an analysis is hence fixed. If the classification is performed so that the test statistic is maximized, this will yield statistics with nonstandard distributions.[22,23] Evaluation of these distributions may then allow evaluation of significance without the need for simulation. An extension of this approach to deal with the situation we describe here (where the number of cases is gradually increased according to the quantitative covariate) would be an interesting area for further study. For nearly all practical purposes, the simulation-based procedure we describe for establishing statistical significance would be tractable.

The SA method may be useful in other scenarios. One further application would include the use of the procedure with affected sib pairs from linkage studies. If only one sib is to be used in an association study, then such individuals may benefit from being sequentially added on the basis of their (pairwise) identity by descent (IBD) proportion. We are currently investigating this application of SA further.

Acknowledgements

We thank the Higher Education Funding Council for Wales, the UK Medical Research Council and the Wellcome Trust for providing financial support. We are indebted to all the individuals who have participated in this research. We thank the anonymous reviewers for their helpful comments on a previous version of this manuscript.

References

1 Xu J, Dimitrov L, Chang BL *et al*: A combined genome-wide linkage scan of 1,233 families for prostate cancer susceptibility genes conducted by the international consortium for prostate cancer genetics. *Am J Hum Genet* 2005; **77**: 219–229.
2 Baron M: Manic-depression genes and the new millennium: poised for discovery. *Mol Psychiatr* 2002; **7**: 342–358.
3 Funalot B, Varenne O, Mas JL: A call for accurate phenotype definition in the study of complex disorders. *Nat Genet* 2004; **36**: 3.
4 Silverman EK, Mosley JD, Rao DC *et al*: Linkage analysis of alpha 1-antitrypsin deficiency: lessons for complex diseases. *Hum Hered* 2001; **52**: 223–232.
5 Craddock N, O'Donovan MC, Owen MJ: The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet* 2005; **42**: 193–204.
6 Craddock N, Jones I, Kirov G, Jones L: The bipolar affective disorder dimension scale (BADDS) – a dimensional scale for rating lifetime psychopathology in bipolar spectrum disorders. *BMC Psychiatry* 2004; **4**: 19.
7 Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M: Ordered subset analysis in genetic linkage mapping of complex traits. *Genet Epidemiol* 2004; **27**: 53–63.
8 Kirkwood BR, Sterne JAC: *Essential Medical Statistics*. Oxford, UK: Blackwell Science, 2003.
9 Efron B: *The Jacknife, the Bootstrap and Other Resampling Plans*. Philadephia: Society for Industrial and Applied Mathematics, 1982.
10 R Development Core Team: *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2004, ISBN 3-900051-00-3.
11 Li Y, Nowotny P, Holmans P *et al*: Association of late-onset Alzheimer's disease with genetic variation in multiple members of the GAPD gene family. *Proc Natl Acad Sci USA* 2004; **101**: 15688–15693.
12 Cummings JL, Vinters HV, Cole GM, Khachaturian ZS: Alzheimer's disease – etiologies, pathophysiology, cognitive reserve, and treatment opportunities. *Neurology* 1998; **51**: S2–S17.
13 Rocchi A, Pellegrini S, Siciliano G, Murri L: Causative and susceptibility genes for Alzheimer's disease: a review. *Brain Res Bull* 2003; **61**: 1–24.
14 Raybould R, Green EK, MacGregor S *et al*: Bipolar disorder and polymorphisms in the dysbindin gene (DTNBP1). *Biol Psychiatry* 2005; **57**: 696–701.
15 Macgregor S, Jones I, Segurado R *et al*: Univariate and multivariate qtl linkage analysis of bipolar affective disorder dimension

scale (BADDS) scores in bipolar disorder. *Am J Med Genet B* 2004; **130B**: 30.

16 O'Mahony E, Corvin A, O'Connell R *et al*: Sibling pairs with affective disorders: resemblance of demographic and clinical features. *Psychol Med* 2002; **32**: 55–61.

17 Straub RE, Jiang YX, MacLean CJ *et al*: Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet* 2002; **71**: 337–348.

18 Williams NM, Preece A, Mortis DW *et al*: Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin binding protein gene (DTNBP1). *Arch Gen Psychiatry* 2004; **61**: 336–344.

19 Klein JP, Moeschberger ML: Survival analysis. *Techniques for Censored and Truncated Data*. New York: Springer, 2004.

20 Dudbridge F: Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 2003; **25**: 115–121.

21 BHF Family Heart Study Research Group: A genomewide linkage study of 1,933 families affected by premature coronary artery disease: The British Heart Foundation (BHF) Family Heart Study. *Am J Hum Genet* 2005; **77**: 1011–1020.

22 Miller R, Siegmund D: Maximally selected chi square statistics. *Biometrics* 1982; **38**: 1011–1016.

23 Koziol JA: On maximally selected chi-square statistics. *Biometrics* 1991; **47**: 1557–1561.