PBAT: A comprehensive software package for genome-wide association analysis of complex family-based studies

Kristel Van Steen[1] *and Christoph Lange*[1,2,*]

[1] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[2] Harvard Medical School, Channing Laboratory, Boston, MA 02115, USA
* *Correspondence to*: Tel: +1 617 432 4919; Fax: +1 617 432 5619; E-mail: clange@hsph.harvard.edu

Date received (in revised form): 7th December 2004

Abstract

The PBAT software package (v2.5) provides a unique set of tools for complex family-based association analysis at a genome-wide level. PBAT can handle nuclear families with missing parental genotypes, extended pedigrees with missing genotypic information, analysis of single nucleotide polymorphisms (SNPs), haplotype analysis, quantitative traits, multivariate/longitudinal data and time to onset phenotypes. The data analysis can be adjusted for covariates and gene/environment interactions. Haplotype-based features include sliding windows and the reconstruction of the haplotypes of the probands. PBAT's screening tools allow the user successfully to handle the multiple comparisons problem at a genome-wide level, even for 100,000 SNPs and more. Moreover, PBAT is computationally fast. A genome scan of 300,000 SNPs in 2,000 trios takes 4 central processing unit (CPU)-days. PBAT is available for Linux, Sun Solaris and Windows XP.

Keywords: *association analysis, extended pedigrees, genome-wide screening, quantitative and qualitative traits, haplotypes*

Genetic association studies take advantage of the fact that we can measure genotypes directly via either protein electrophoretic or molecular genetic methods. The goal is to explain the variation in the disease trait of interest using an individual's genotype as a genetic marker. There are two basic types of study design that are used in genetic association analysis: standard (population-based, case-control or cohort) and family-based. Analytical methods appropriate for these two designs are quite different. The family-based design is attractive for many reasons. For one, the design protects against a finding of spurious association, due to population admixture or stratification. The reason for robustness is that the analysis uses parental genotypes to determine the distribution of the test statistic. The analysis cannot be biased by admixture or stratification because the case and control alleles are drawn from the same subjects; therefore, they have the same genetic background. The other key advantage of family-based studies is the way the multiple testing problem can be handled. Using the conditional mean model approach,[1–3] the data are first analysed in a 'screening step'. The analysis of the screening step does not bias the significance level of subsequently computed tests. In this screening step, the scientist can look at all possible associations between the markers and traits and select a subset of 'promising' marker–trait combinations — typically five combinations.[3] Only the selected subset is then put forward to the hypothesis-testing step.

A general paradigm for testing the association between a response variable (disease trait) and a predictor (genotype as a marker) is a regression analysis, since this can accommodate all types of outcomes and all types of predictors. Although regression analysis has many advantages and is widely used in epidemiological investigations, it does require specifying a model for how the trait depends upon the genotype. If the model is incorrect, the power may be reduced. Depending upon study design and analysis, there may also be consequences for the validity. Cordell and Clayton[4] have described a unified approach to performing genetic association analysis with nuclear families (or case/control data) in a regression context. Case–parent trios are analysed via conditional logistic regression using the case and three pseudo-controls derived from the untransmitted parental alleles. The beauty of the method is that it can be performed using standard statistical software and that additional effects, such as parent-of-origin, effects can be included. The major drawback is that, to date, the technique has not been adapted to include extended pedigrees without splitting them up into nuclear families.

A large number of computer programs are available for family-based association tests, including AFBAC,[5] QTDT,[6]

FBAT,[7–11] TRANSMIT[12] and PDT.[13] These software packages primarily focus on the computation of various test statistics, whereas the PBAT software package also exhibits pre- and post-analysis features. The PBAT software can be downloaded from http://www.biostat.harvard.edu/~clange/default.htm.

PBAT is an interactive software package that provides tools for the design and data analysis of family-based association studies. It is available for Windows XP, Linux and UNIX operating systems. The newest version of PBAT (v2.5) includes many features that were not available in earlier versions,[14] such as haplotype analysis tools that can be invoked using batch mode or user interface, more flexible specifications in power calculations and allowance for discrete trait distribution when applicable. In particular, PBAT incorporates the features of the family-based tests of association (FBAT) package (http://www.biostat.harvard.edu/fbat/fbat.htm) but provides many additional options for designing association/linkage studies and analysing data with multiple continuous traits. Perhaps the most striking feature, which gives PBAT a unique advantage over most available software in the field, is its implementation of the screening techniques — that is, the conditional mean model approach[1,2] — that allow the user to handle the multiple comparison problem at a genome-wide level.[3] Further advantages of PBAT are the analytical power and sample size calculations for family-based association tests.[15,16] PBAT is especially well suited for quantitative traits while possibly accounting for important predictors.

The cornerstone of the package is the unified approach to FBAT, introduced by Rabinowitz and Laird[17] and Laird *et al.*[10]. FBAT builds on the original Transmission Disequilibrium Test (TDT) method,[18] in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring. It has been generalised so that tests of different genetic models, tests of different sampling designs, tests involving different disease phenotypes, tests with missing parents and tests of different null hypotheses are all in the same framework. In particular, the FBAT statistic is based on a linear combination of offspring genotypes and traits:

$$\mathrm{FBAT} = (S - E[S])/V^{1/2}, \quad S = \Sigma_{ij} T_{ij} * X_{ij} \qquad (1)$$

where $V = \mathrm{Var}(S)$ and T_{ij} represents the coded phenotype (ie the phenotype adjusted for any covariates) of the j-th offspring in family i. X_{ij} denotes the offspring's coded genotype at the locus being tested. It depends on the genetic model under consideration.

The expected distribution is derived using Mendel's law of segregation and conditioning on the sufficient statistics for any nuisance parameters under the null hypothesis, the null hypothesis being 'no linkage and no association' or 'no association, in the presence of linkage'.

PBAT provides methods for a wide range of situations that arise in family-based association studies using FBAT statistics. More specifically, there are two main components: tools for the planning of family-based association studies and data analysis tools. In terms of study planning, PBAT computes the power for study designs that consist of different family types with varying numbers of offspring, under different ascertainment conditions and allowing for missing parental genotypes. The data analysis tools available in PBAT provide options to test linkage or association in the presence of linkage, using (bi-allelic or multi-allelic) marker or haplotype data, single or multiple traits (eg measurements recorded repeatedly over time) that may be quantitative, qualitative or time-to-onset, with nuclear families as well as extended pedigrees. PBAT easily handles covariates and gene/covariate interactions in all computed FBAT statistics. Furthermore, PBAT can also be used for post-study power calculations and construction of the most powerful test statistic. For situations in which multiple traits and markers are given, PBAT's screening tools reduce the large pool of traits and markers and select the most promising combinations in terms of the FBAT statistic.

Using PBAT's screening tools the present authors have shown that genome-wide association studies using families are realisable in terms of data analysis.[3] The key concept of the implemented screening techniques is the conditional mean model approach,[1,2] for which the data space is partitioned into two independent testing sets. This allows one to control the type I error rates and to overcome one of the most important statistical hurdles when analysing genome-wide association studies with thousands of markers: the multiple comparison problem. The screening technique maintains its protective character for extended datasets with a few hundred thousand SNPs. It should be noted that, in general, adding more SNPs comes at the cost of power loss when corrections for multiple testing need to be applied (eg Bonferroni-type corrections to control type I error). These screening methods are hardly affected by adding 'non-causal' SNPs. In addition, they are robust against effects of population stratification and admixture, since the final decision in the screening process is based on FBATs, which guard against these confounding factors. Finally, PBAT's screening tools are most successful in detecting common disease susceptibility loci. This is particularly attractive in the light of the HapMap project,[19] which aims to describe the common patterns of genetic variation in humans.

The problem of detecting rare disease-associated SNPs remains; however, this is a general problem rather than a problem specifically related to the screening techniques of PBAT. Applying the authors' screening tools using the haplotype features of PBAT (eg using sliding windows acknowledging the linkage disequilibrium structures present in the data) may be more beneficial. This is work in progress. TRANSMIT[12] is another program for transmission disequilibrium testing that uses marker haplotypes based on several closely linked markers. By contrast with PBAT, however, TRANSMIT leads to elevated false-positive rates in the presence of population admixture and does not handle

quantitative traits.[20] Moreover, it has no built-in functions for performing screening on a genome-wide level.

PBAT's data analysis tools have been extensively validated. These include the data analysis tools using univariate and multivariate traits,[21] multivariate/longitudinal FBAT models,[22] time-to-onset traits (Su; personal communication), haplotype analysis (Randolph; personal communication) and genomic screening.[3] PBAT is under constant development. Future developments include refined screening tools and guidelines that apply to haplotype-based genomic screening, power calculations for haplotype analysis and further effort towards a PBAT compendium of commands and an extensive documentation for its users.

References

1. Lange, C., DeMeo, D.L., Silverman, E. *et al.* (2003), 'Using the noninformative families in family-based association tests: A powerful new testing strategy', *Am. J. Hum. Genet.* Vol. 73, pp. 801–811.
2. Lange, C., Lyon, H., DeMeo, D.L. *et al.* (2003), 'A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies', *Hum. Hered.* Vol. 56, pp. 10–17.
3. Van Steen, K., McQueen, M.B., Herbert, A. *et al.* (2005), 'Genomic screening in family based association testing for quantitative traits', *Nat. Genet.* (under review).
4. Cordell, H.J. and Clayton, D.G. (2002), 'A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in Type 1 diabetes', *Am. J. Hum. Genet.* Vol. 70, pp. 124–141.
5. Thomson, G. (1995), 'Mapping disease genes: Family-based association studies', *Am. J. Hum. Genet.* Vol. 57, pp. 487–498.
6. Abecasis, G.R., Cardon, L.R. and Cookson, W.O.C. (2000), 'A general test of association for quantitative traits in nuclear families', *Am. J. Hum. Genet.* Vol. 66, pp. 279–292.
7. Horvath, S. and Laird, N. (1998), 'Discordant sibship test for disequilibrium/transmission: No need for parental data', *Am. J. Hum. Genet.* Vol. 63, pp. 1886–1897.
8. Horvath, S., Xin, X. and Laird, N.M. (2000), 'The family based association test method: Computing means and variances for general statistics', *Technical Report*, http://www.biostat.harvard.edu/fbat/fbattechreport.ps.
9. Horvath, S., Xu, X. and Laird, N.M. (2001), 'The family based association test method: Strategies for studying general genotype-phenotype associations', *Eur. J. Hum. Genet.* Vol. 9, pp. 301–306.
10. Laird, N.M., Horvath, S. and Xu, X. (2000), 'Implementing a unified approach to family-based tests of association', *Genet. Epidemiol.* Vol. 19(1), pp. S36–S42.
11. Lake, S., Blacker, D. and Laird, N.M. (2000), 'Family-based tests of association in the presence of linkage', *Am. J. Hum. Genet.* Vol. 67, pp. 1515–1525.
12. Clayton, D. (1999), 'A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission', *Am. J. Hum. Genet.* Vol. 65, pp. 1170–1177.
13. Martin, E.R., Monks, S.A., Warren, L.L. *et al.* (2000), 'A test for linkage and association in general pedigrees: the pedigree disequilibrium test (PDT)', *Am. J. Hum. Genet.* Vol. 67, pp. 146–154.
14. Lange, C., DeMeo, D.L., Silverman, E.K. *et al.* (2004), 'PBAT: Tools for family-based association studies', *Am. J. Hum. Genet.* Vol. 74, pp. 367–369.
15. Lange, C. and Laird, N.M. (2002), 'Power calculations for a general class of family-based association tests: Dichotomous traits', *Am. J. Hum. Genet.* Vol. 67, pp. 575–584.
16. Lange, C., DeMeo, D. and Laird, N.M. (2002), 'Power calculations for a general class of family-based association tests: Quantitative traits', *Am. J. Hum. Genet.* Vol. 71, pp. 575–584.
17. Rabinowitz, D. and Laird, N.M. (2000), 'A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information', *Hum. Hered.* Vol. 50, pp. 227–233.
18. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993), 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am. J. Hum. Genet.* Vol. 65, pp. 578–580.
19. The International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
20. Horvath, S., Xu, X., Lake, S.L. *et al.* (2004), 'Family-based tests for associating haplotypes with general phenotype data: Application to asthma genetics', *Genet. Epidemiol.* Vol. 26, pp. 61–69.
21. DeMeo, D.L., Lange, C., Silverman, E.K. *et al.* (2002), 'Univariate and multivariate family-based association analysis of the IL-13 ARG130GLN polymorphism in the Childhood Asthma Management Program', *Genet. Epidemiol.* Vol. 23, pp. 335–348.
22. Lange, C., Van Steen, K., Andrew, T. *et al.* (2004), 'A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects', *Stat. Appl. Genet. Mol. Biol.* Vol. 1(1), Article 17, http://www.bepress.com/sagmb/vol3/iss1/art17.