BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 1 2009, pages 132–133
doi:10.1093/bioinformatics/btn563

*Genetics and population analysis*

# A multivariate test of association

Manuel A. R. Ferreira[1,2,3,4,5,*] and Shaun M. Purcell[1,2,3,4,6,7]

[1]Department of Psychiatry, Massachusetts General Hospital, [2]Department of Psychiatry, Harvard Medical School, [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, [4]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA, [5]Genetic Epidemiology, Queensland Institute of Medical Research, QLD, Australia, [6]Broad Institute of Harvard and MIT, Cambridge and [7]Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

Received on August 3, 2008; revised on October 22, 2008; accepted on October 24, 2008

Advance Access publication November 19, 2008

Associate Editor: Alex Bateman

**ABSTRACT**

**Summary:** Although genetic association studies often test multiple, related phenotypes, few formal multivariate tests of association are available. We describe a test of association that can be efficiently applied to large population-based designs.

**Availability:** A C++ implementation can be obtained from the authors.

**Contact:** manuel.ferreira@qimr.edu.au

**Supplementary information:** Supplementary figures are available at *Bioinformatics* online.

Genetic association studies often test multiple traits. For example, for many diseases, such as asthma or attention deficit hyperactivity disorder (ADHD), investigators routinely measure multiple endophenotypes that are thought to be more proximal to the biological etiology of the clinical disorder. The expectation that often underlies the analysis of multiple traits is that this strategy can identify not only trait-specific quantitative trait loci (QTL), but also those shared between correlated traits.

Analysis of such multivariate datasets typically consists of testing each trait individually and then informally comparing the evidence for association at a particular locus across traits. This approach, however, has two major caveats: first, if unaccounted for, multiple trait testing increases the experiment-wise false-positive rate. Second, it ignores the extra information provided by the cross-trait covariance intrinsic to multivariate datasets.

Although these limitations can be alleviated to some extent by pre- or post-analytic strategies, such as principal components analysis or permutation testing, these can be inefficient and/or computationally intensive when a large number of traits or loci are under investigation. Here, we describe an efficient multivariate test of association for population-based designs that we have implemented in PLINK (Purcell *et al*., 2007).

Consider a sample of $n$ unrelated individuals with data for two sets of variables, a bi-allelic marker (set 1, with one variable) and $k$ traits (set 2). We use canonical correlation analysis (CCA), which is a multivariate generalization of the Pearson product-moment correlation (Hotelling, 1936), to measure the association between the two sets of variables. Specifically, CCA extracts the linear combination of traits that explain the largest possible amount of the covariation between the marker and all traits. Although this approach is most appropriate for the analysis of normally distributed traits, as we show below, it shows good performance even when considering non-normal or disease traits.

Prior to the actual CCA, the marker is coded according to an allelic dosage scheme that can incorporate dominance; our approach can also be extended to the analysis of multiple markers by expanding the first set of variables to include more than one marker. Missing phenotype data are handled either by case-wise deletion (if data are missing above a pre-defined per-individual missingness threshold) or mean imputation (i.e. a missing phenotype is replaced by the sample mean). The test is based on Wilk's lambda ($\lambda$) and the corresponding $F$-approximation, both simplified to the situation where one of the sets contains only one variable (the marker). Specifically, $\lambda = 1-\hat{\rho}^2$, where $\hat{\rho}$ is the canonical correlation between the marker and the $k$ traits, calculated as the square root of the eigenvalue of $S_{11}^{-1/2} \cdot S_{12} \cdot S_{22}^{-1} \cdot S_{21} \cdot S_{11}^{-1/2}$. In the latter, $S_{11}$ is the marker variance, $S_{22}$ the $k \times k$ trait covariance matrix, while $S_{12}$ and $S_{21}$ are the $1 \times k$ (or $k \times 1$) covariance matrices between the marker and the $k$ traits. The simplified $F$-approximation is:

$$F_{(k,n-k-1)} = \left[(1-\lambda)/\lambda\right] \cdot \left[(n-k-1)/k\right].$$

The interpretation of a significant multivariate test is aided by the inspection of the weights attributed by the CCA to each phenotype.

To investigate the performance of the proposed method, we simulated data for five quantitative traits (60% heritability each) and one bi-allelic locus (20% minor allele frequency). The QTL explained 0.5% of the total variance of 0 (to assess type-I error rate), 1, 2, 3, 4 or all 5 traits. We considered residual cross-trait correlations (i.e. excluding the QTL effect) of 0.2, 0.4 or 0.6 (with the same sign as the QTL-induced correlation), and simulated data for 600 unrelated individuals. We compared the power of the multivariate approach against two univariate strategies: first, we tested each trait individually using linear regression, selected the most significant test and correct this for multiple testing through the analysis of 100 permuted datasets. Each dataset was generated by randomly permuting the genotypes between individuals, thus preserving the original cross-trait correlations. For the second strategy, a simple Bonferroni correction based on the number of traits analyzed was applied to the most significant univariate test. Power and type-I error (nominal $\alpha = 0.05$) for each model were based on the analysis of

[*]To whom correspondence should be addressed.

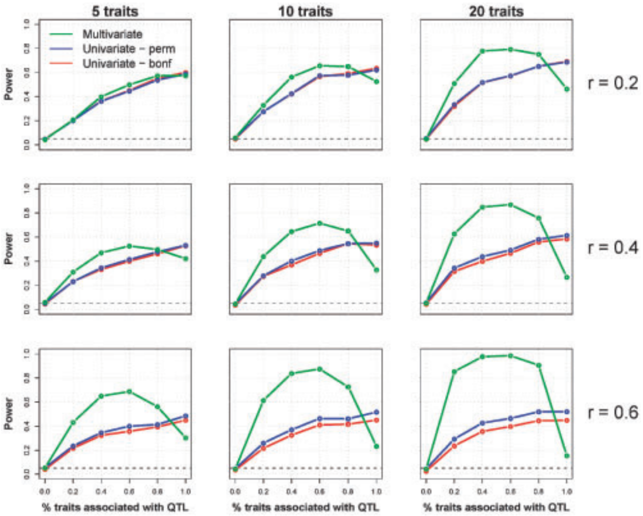

**Fig. 1.** Performance of the multivariate test of association. Type-I error and power are also shown for two univariate strategies which correct for multiple testing through permutation or simple Bonferroni correction. *r* indicates, residual trait correlation. See text for details.

1000 and 5000 replicates, respectively. This same procedure was used to test the performance when analyzing 10 and 20 traits.

The simulation results are shown in Figure 1. When considering five traits with a modest residual cross-trait correlation ($r = 0.2$), the power of the multivariate test was comparable to both univariate strategies considered, with the advantage that no permutation testing or Bonferroni adjustment was required to correct for multiple testing. As the number of traits, or the residual correlation between traits, increased, the power of multivariate test improved, consistently outperforming the univariate approaches. The exception was for the extreme models where all traits were associated with the QTL; in this case, the power of the multivariate test decreased as the residual correlation between traits increased. This observation is consistent with previous reports (Allison *et al.*, 1998; Amos *et al.*, 2001; Evans and Duffy, 2004; Ferreira *et al.*, 2006) and was specific to the situation where the QTL-induced trait correlation was of the same sign as the correlation induced by residual shared factors (Supplementary Fig. 1).

The multivariate test maintained appropriate type-I error when some or all traits tested were continuous but not normally distributed (Supplementary Fig. 2) or were measured on a discrete scale (Supplementary Fig. 3).

Finally, we also extended this approach to the analysis of family-based data. Briefly, prior to CCA, each individual's genotype is partitioned into the orthogonal between- (B) and within-family (W) components (Fulker *et al.*, 1999). We then perform CCA using the *k* traits and either the B (between-family association test), W (within-family association test, which is robust to population stratification effects) or the B + W (total association test) genotype scores, and use an adaptive permutation procedure to account for family structure. Simulations show that when applied to family data, this approach also has appropriate type-I error and improved power when compared to the univariate strategy (Supplementary Fig. 4).

In conclusion, we propose a robust and powerful test of association that can accommodate multiple phenotypes and different study designs. As such, it can be relevant to many genetic association studies of complex traits or diseases.

*Funding*: Sidney Sax Fellowship from the National Health and Medical Research Council of Australia (to M.A.R.F.).

*Conflict of Interest*: none declared.

## REFERENCES

Allison,D.B. *et al.* (1998) Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.*, **63**, 1190–1201.

Amos,C. *et al.* (2001) Comparison of multivariate tests for genetic linkage. *Hum. Hered.*, **51**, 133–144.

Evans,D.M. and Duffy,D.L. (2004) A simulation study concerning the effect of varying the residual phenotypic correlation on the power of bivariate quantitative trait loci linkage analysis. *Behav. Genet.*, **34**, 135–141.

Ferreira,M.A. *et al.* (2006) A simple method to localise pleiotropic susceptibility loci using univariate linkage analyses of correlated traits. *Eur. J. Hum. Genet.*, **14**, 953–962.

Fulker,D.W. *et al.* (1999) Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.*, **64**, 259–267.

Hotelling,H. (1936) Relations between two sets of variables. *Biometrika*, **28**, 321–377.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.