A gene-based test of association using canonical correlation analysis

Clara S. Tang and Manuel A. R. Ferreira*

Queensland Institute of Medical Research, Brisbane, QLD 4029, Australia

ABSTRACT
Motivation: Canonical correlation analysis (CCA) measures the association between two sets of multidimensional variables. We reasoned that CCA could provide an efficient and powerful approach for both univariate and multivariate gene-based tests of association without the need for permutation testing.

Results: Compared with a commonly used permutation-based approach, CCA (i) is faster; (ii) has appropriate type-I error rate for normally distributed quantitative traits; (iii) provides comparable power for small to medium-sized genes (<100 kb); (iv) provides greater power when the causal variants are uncommon; (v) provides considerably less power for larger genes (>100 kb) when the causal variants have a broad minor allele frequency (MAF) spectrum. Application to a GWAS of leukocyte levels identified variants with a broad MAF spectrum. However, CCA analysis showed that the causal variants were likely to be located in the promoter region of the gene, thus indicating the importance of considering the correlation between SNPs when testing for association.


Supplementary information: Supplementary material is available at Bioinformatics online.

Received on October 6, 2011; revised on January 3, 2012; accepted on January 20, 2012

1 INTRODUCTION
Over the past few years, genome-wide association studies (GWASs) have been successful in uncovering common genetic variants underlying complex traits or diseases. Considerable efforts have been made to replicate the most associated marker in a given GWAS locus, but there is growing evidence that some loci harbor additional independent associations (Moffatt et al., 2010; Saccone et al., 2009; Zheng et al., 2007). This phenomenon, which may be widespread, has been explored through gene-based association approaches (Neale and Sham, 2004) to facilitate the identification of genes with multiple variable associations with weak effects.

Current gene-based methods typically require permutation or simulation testing to account for the correlation between SNPs as well as gene size, among which VEGAS (Liu et al., 2010) and PLINK set-based tests (Purcell et al., 2007) are most frequently used. Both tests combine results from single-SNP analyses but differ in how an appropriate null distribution is obtained. VEGAS relies on simulations of linkage disequilibrium (LD) structure from a reference dataset, such as the HapMap, whereas PLINK set-based tests resort to permutation testing. In general, these are flexible approaches that can accommodate many different analytical strategies, but can be computationally intensive with increasing number of SNPs tested and when the level of significance is high.

Previously, Ferreira and Purcell (2009) proposed a multivariate test of association based on canonical correlation analysis (CCA) to simultaneously test the association between a single SNP and multiple phenotypes. CCA, first described by Harold Hotelling (1936), measures the association between two sets of multidimensional variables by maximizing the correlation between their linear combinations. We reasoned that CCA could also provide an efficient and powerful approach for a gene-based test of association by intrinsically taking into account the correlation structure between SNPs without the need for permutation testing.

In this study, we specifically tested the performance of CCA for the gene-based scenarios (i) and (ii) and (vi) and (vii).

To implement the CCA test, each SNP in the target region is first coded according to the allelic dosage (0, 1 and 2). Next, to remove any multicollinearity between SNPs, we perform a two-stage LD pruning step: first, we remove SNPs that are in high pairwise LD (r^2 > 0.8) with other markers; second, to identify and remove high correlations between linear combinations of SNPs, we use variance inflation factor (VIF) analysis to iteratively exclude independent SNPs that have a VIF > 2 with other markers.

When p > 1, the same pruning procedure is applied to the phenotype-set to remove any strongly correlated phenotypes from the analysis. We tested different LD and VIF thresholds but found that 0.8 and 2, respectively, resulted in appropriate type-I error and optimal power.

Finally, for a sample of n unrelated individuals with data for q SNPs and p phenotypes, the j = min(q,p) canonical correlations $\rho_j$ are then calculated as the square root of the j eigenvalues of the canonical correlation matrix.
There is growing demand for powerful and efficient single-trait gene-based tests of association. We performed extensive simulations to assess the performance of CCA. To test the performance of CCA, we simulated SNP data for a hypothetical gene with the coalescent-based simulator GENOME (Li et al., 2007). We assumed a mutation rate of \(10^{-8}\) per generation per base pair and an effective population size of 10,000. A study population of 2000 individuals was formed by (i) simulating 1000 haplotypes; (ii) pairing these to form a pool of 500 diploid genomes; and (iii) randomly drawing with replacement 2000 diploid genotypes from that pool. SNPs with a minor allele frequency (MAF) < 0.01 were excluded from analysis. For some of the models tested (see below), the simulated gene included multiple independent SNPs (or QTL, with \(k \geq 1\)) contributing to phenotypic variation. To achieve this, \(k\)-length recombination hotspots were created by simulating SNP data for \(k\) regions independently, which were then appended to form the complete gene.

### 2.2 Coalescent-based simulation of SNP data

To test the performance of CCA, we simulated SNP data for a hypothetical gene with the coalescent-based simulator GENOME (Li et al., 2007). We assumed a mutation rate of \(10^{-8}\) per generation per base pair and an effective population size of 10,000. A study population of 2000 individuals was formed by (i) simulating 1000 haplotypes; (ii) pairing these to form a pool of 500 diploid genomes; and (iii) randomly drawing with replacement 2000 diploid genotypes from that pool. SNPs with a minor allele frequency (MAF) < 0.01 were excluded from analysis. For some of the models tested (see below), the simulated gene included multiple independent SNPs (or QTL, with \(k \geq 1\)) contributing to phenotypic variation. To achieve this, \(k\)-length recombination hotspots were created by simulating SNP data for \(k\) regions independently, which were then appended to form the complete gene.

#### 2.3.3 Gene length

We simulated data for five gene lengths: 20, 25, 50, 100 and 500 kb. We set \(k=5\) and \(h^2=0.9\%\).

#### 2.3.4 QTL allele frequency

Instead of randomly selecting SNPs to be treated as QTL, in these analyses we imposed a MAF constraint such that only SNPs with a MAF within a narrow range (1–5%, 5–10%, 10–20% or 20–50%) were selected as QTL. We also varied gene length from 20 to 500 kb, but kept \(k=5\) and \(h^2=0.9\%\).

#### 2.3.5 Allele frequency of SNPs included for analysis

Lastly, in addition to controlling the MAF of the simulated QTL, we also restricted the MAF of the SNPs included for analysis within two ranges, uncommon (1–5%) and relatively common (20–50%) SNPs. In these analyses, gene length was 50 kb, \(k=5\) and \(h^2=0.9\%\).

For each model, 1000 simulations (25 000 for type-I error rate) were performed. Power (or type-I error rate) was then estimated as the proportion of simulations with a \(P\)-value significant for \(a=0.01\) (unless otherwise noted). The performance of the CCA test was compared against two standard permutation-based strategies implemented in PLINK (Purcell et al., 2007).

One such strategy tests each SNP in a gene for association with the phenotype, computes the average chi-square statistic across all SNPs, and then assesses its significance through permutation. Alternatively, the best-SNP test considers only the maximum chi-square statistic across all SNPs tested and determines its significance through permutation. We also analyzed each simulated dataset with GWiS (Huang et al., 2011), which uses greedy Bayesian model selection to identify independent effects within a gene and estimates overall significance through permutation. We used the default GWiS analysis parameters except the maximum number of permutations (set to 1000). Both PLINK tests and GWiS were conducted on the full dataset, i.e. prior to the pruning step.

### 2.4 Gene-based analysis of multiple quantitative traits

To explore the performance of CCA as a multi-phenotype gene-based test of association, we simulated SNP data for a 60 kb gene with \(k=3\) and \(h^2=0.9\%\), as described above. Next, we simulated five normally distributed traits under five models. In Model 1, each of the three independent QTL individually explained 0.3% (0.9%) of the variance of trait 1, i.e. the QTL had no impact on the variance of the other four traits. In Models 2–5, the three QTL contributed to the variance of a progressively larger number of traits: 2, 3, 4 and all 5 traits, respectively. In all five models, shared sources of variation between traits other than the QTL were also simulated so that the residual phenotypic correlation between pairs of traits was \(\sim 0.4\). We also considered a null model with \(h^2=0\%\) to assess type-I error rate. We compared the power of the CCA test with a permutation-based approach that estimates the empirical significance of the average chi-square statistic obtained when testing all SNPs for association with all traits.

### 2.5 Gene-based analysis of a single disease trait

We also investigated if CCA could provide a valid gene-based test for the analysis of a single disease trait. For this purpose, we assumed a liability-threshold model and first simulated SNP data and a normally distributed trait as described above for 2000 individuals, assuming a disease prevalence of 1.5 or 10%. For a disease prevalence \(p\), individuals with a trait value above \(Q(d)\) [e.g. for \(d=0.05\), \(Q(0.05)=1.64\)] were considered as cases, and as controls otherwise. The CCA test was then applied as described above. In these analyses, gene length was 50kb, \(k=5\) and \(h^2=0.9\%\). Type-I error rate was assessed by setting \(h^2=0\%\).

### 2.6 Application to a GWAS of white blood cell traits

To illustrate the applicability of CCA as a gene-based test of association, we considered six white blood cell (WBC) traits measured in 1061 unrelated individuals of European ancestry, including total WBC, neutrophil, lymphocyte, monocyte, eosinophil and basophil counts. This dataset is a...
subset of the family-based Australian cohort analyzed by Ferreira et al. (2009). Each trait was normalized and adjusted for age and sex effects prior to the analysis. Univariate outlier observations (6 SD above the mean) were excluded from analysis. Data were available for 2.3 million autosomal SNPs with MAF > 1%, including directly genotyped (obtained with Illumina 610 K arrays) and imputed (based on HapMap 2) SNPs. We tested each individual trait (univariate gene-based analyses) or all traits simultaneously (multivariate gene-based analysis) for association with 17 470 genes. Gene boundaries were defined by considering the start and end of the largest isoform for each gene according to the annotation in the UCSC Genome Browser (hg18, NCBI build 36). We added a 15 kb buffer upstream and downstream as most cis-regulatory variants fall within 15 kb of the respective gene (Pickrell et al., 2010).

3 RESULTS

3.1 Simulation approach
To test the performance of CCA as a gene-based test of association, we simulated SNP data for 2000 individuals under a coalescent model and restricted our analyses to SNPs with a MAF ≥ 1% (Supplementary Fig. S1). Data were simulated for a hypothetical gene ranging in length from 20 kb (17 SNPs on average) to 500 kb (average of 463 SNPs). A single or multiple SNPs in the gene were then selected randomly as causal variants (i.e. QTL) and a normally distributed trait simulated for a range of models as described in the ‘Methods’ section.

3.2 Gene-based analysis of a single quantitative trait
We studied in detail the performance of CCA when applied to model selection to identify independent effects within a gene and is implemented in GWiS (Huang et al., 2011). We tested five different factors that were likely to affect the power of the CCA test: (i) proportion of the phenotypic variance explained by the gene; (ii) number of independent QTL; (iii) gene length; (iv) QTL allele frequency; and (v) allele frequency of SNPs included for analysis.

3.2.1 Proportion of phenotypic variance explained by the gene ($h^2$) We simulated data for a 50 kb gene with three independent QTL and varied $h^2$ from 0.6% to 1.8%. All four tests provided comparable power, which improved with increasing effect size (Fig. 1A).

3.2.2 Number of independent QTL ($k$) As expected, increasing the number of independent QTL—while maintaining $h^2$ constant—led to a steady power loss for the best-SNP test (Fig. 1B), which only takes into account the result for the most associated SNP in the gene. Similar results were obtained with GWiS. In contrast, the power of the CCA and all-SNP approaches was comparable and largely unaffected by the number of independent QTL.

3.2.3 Gene length ($l$) Permutation- or simulation-based methods can become computationally intensive for large genes with hundreds of SNPs genotyped (see Supplementary Table S2 for a comparison of running times). A test that is both efficient and powerful even when analyzing large genes is therefore desirable and so next we investigated the impact of gene size on the performance of the CCA test. For smaller genes ($\leq 50$ kb), the power of the CCA and all-SNP tests was comparable, and both outperformed the best-SNP and GWiS tests (Fig. 1C). However, the power of the CCA test decreased rapidly with increasing gene size, becoming comparatively less powerful than the all-SNP test for large genes ($\geq 100$ kb). The power of the all-SNP, best-SNP and GWiS tests also decreased with increasing gene size, but less markedly.

3.2.4 QTL allele frequency We also assessed the impact of the QTL MAF on the power of the CCA test. Power was comparable

![Fig. 1. Power of the CCA gene-based test when analysing a single quantitative trait as a function of the proportion of phenotypic variance explained by the gene ($h^2$), the number of independent QTL ($k$) and gene length ($l$). (A) Impact of gene effect size: $h^2$ varied between 0.6% and 1.8%, while $k = 3$ and $l = 50$ kb. (B) Impact of the number of independent QTL: $k$ varied between 3 and 7, while $h^2 = 0.9%$ and $l = 50$ kb. (C) Impact of gene length: $l$ varied between 20 kb and 500 kb, while $k = 5$ and $h^2 = 0.9%$. The performance of the CCA test was compared against two permutation-based approaches implemented in PLINK (best-SNP and all-SNP tests) and GWiS, a recently described gene-based test of association (Huang et al., 2011).]
3.3 Gene-based analysis of multiple quantitative traits

CCA provides a useful framework to simultaneously measure the association between multiple SNPs and multiple traits in a single test and so we were interested in determining how well it would perform when compared to a permutation-based approach. When the three independent QTL simulated in a 60kb gene influenced the variation of only a subset of the five simulated traits, the CCA test provided improved power when compared with the average chi-square test (Supplementary Fig. S4). However, power of the CCA test was considerably lower when the three QTL influenced all five traits, an observation that is consistent with previous reports (Allison et al., 1998; Amos et al., 2001; Evans and Duffey, 2004; Ferreira and Purcell, 2009).

3.4 Gene-based analysis of a single disease trait

Many GWASs focus on complex diseases rather than quantitative traits. Therefore, we were interested in investigating if CCA could also provide a valid gene-based test when analyzing a single disease trait. We simulated data for a 50 kb gene with five independent QTL and a disease trait under a liability-threshold model, assuming a disease prevalence of 1, 5 or 10%. The type-I error rate of the CCA test was again close to the expected nominal level (Supplementary Table S4). For a disease prevalence of 1%, power of the CCA was comparable with that of the all-SNP test (Supplementary Table S5). However, when we simulated data based on a higher disease prevalence (i.e. lower SNP genotype relative risk), power was lower for the CCA test when compared with the all-SNP test.

3.5 Application to a GWAS of WBC counts

We applied the CCA gene-based test to the analysis of six WBC traits measured in 1061 unrelated individuals of European ancestry, including total WBC, neutrophil, lymphocyte, monocyte, eosinophil and basophil counts. We tested each individual trait (i.e. six univariate gene-based analyses) and all traits simultaneously (i.e. a single multivariate gene-based analysis) for association with 17 470 genes.

QQ plots for the six univariate gene-based analyses are provided in Supplementary Fig. S5. No gene-based association exceeded a P = 4.8 × 10^{-7}, a conservative threshold that corrects for the analysis of 17 470 genes and six traits. The most significantly associated genes (P < 5 × 10^{-6}) are listed in Table 1, and included HIST1H4D (total WBC; P = 1.4 × 10^{-5}) and neutrophil counts (P = 3.0 × 10^{-6}); PHKA2A for total WBC (P = 2.2 × 10^{-6}) and lymphocyte counts (P = 4.1 × 10^{-6}); and both C19orf70 (P = 7.8 × 10^{-7}) and SAFB (P = 7.4 × 10^{-7}) for lymphocyte counts.

Closer inspection of these results indicated that all six were driven by at least two independent SNPs nominally associated with the respective phenotype (Supplementary Table S6). For example, of the 42 SNPs located in or within 15 kb of PHKA2A, six were retained for the gene-based analysis after LD-pruning (all with r² < 0.2 with each other), of which four were nominally associated with total WBC counts in single-SNP analyses: rs3809727 (P = 0.0023, MAF = 31%), rs12245600 (P = 0.0105, MAF = 31%), rs11595249 (P = 0.0126, MAF = 34%) and rs17418706 (P = 0.0136, MAF = 13%). Therefore, the region demarcated by this gene contains multiple independent putative QTL for total WBC counts; individually, these have unremarkable associations, but when analyzed jointly the evidence for association with total WBC is considerably greater (CCA gene-based P = 2.2 × 10^{-6}).

We repeated the same six gene-based analyses but restricted these to SNPs with MAF between 1% and 5%, or between 20% and 50%, to search for genes with potential multiple independent uncommon or common variants, respectively. QQ plots for these analyses are...
provided in Supplementary Figs S6 and S7. Although no additional regions were identified with notable gene-based associations, these analyses provided stronger support ($P < 4.8 \times 10^{-12}$) for an association between lymphocyte counts and the chromosome 19 region that includes $C19orf70$ and SAFB (Supplementary Table S7), as well as between neutrophil counts and the histone 1 gene cluster on chromosome 6 (Supplementary Table S8).

Finally, the multivariate CCA gene-based analysis did not identify any new region with a $P < 2.8 \times 10^{-6}$, which corrects for 17 470 tests. However, it provided strong support for SAFB ($P = 2.0 \times 10^{-6}$) and consistent, albeit weaker associations for $HIST1H4D$, $PI4K2A$ and $C19orf70$ (Table 1).

4 DISCUSSION

We performed a series of simulations to test whether CCA could provide a convenient, robust and powerful statistical framework for a gene-based test of association when analyzing a quantitative trait measured in unrelated individuals. Our results suggest that, when compared with a commonly used permutation-based approach that considers all genetic variation assayed in a gene, a CCA gene-based test (i) can be orders of magnitude faster, particularly for larger, highly significant genes; (ii) has appropriate type-I error rate when analyzing a single normally distributed quantitative trait; (iii) provides comparable power for small to medium-sized genes ($<100\,\text{kb}$); (iv) provides greater power when the causal variants are uncommon, irrespective of gene size. On the other hand, the CCA test was considerably less powerful for larger genes ($\geq 100\,\text{kb}$) when the causal variants have a broad MAF spectrum or are relatively common (MAF $> 20\%$). Of note, $22\%$ of the 17 470 genes tested in the WBCs study were larger than 100 kb, which included a $\pm 15\,\text{kb}$ buffer.

We further show that including in the CCA gene-based test all genetic variation assayed in a given gene can be counterproductive when the specific aim of the analysis is to identify genes that harbor exclusively uncommon or common causal variants, but not both. Based on these intuitive results, we suggest that CCA may provide a particularly useful gene-based approach for the separate analysis of either uncommon or common variants.

Gene-based association $P$-values are often used as input for pathway-based association analyses. In this case, CCA may not be the most appropriate test to use because of the large impact of gene size on power. Gene-based methods that have a more comparable performance across different gene lengths (e.g. all-SNP or GWAS tests) may provide a better alternative to ensure that an association with pathways composed of large genes can be as readily detected as with those composed of smaller genes.

Our results suggest that CCA also provides a valid test for the analysis of a single disease trait or multiple quantitative traits. In this case, however, CCA provided little if any gain in power when compared to the all-SNP permutation-based approach. The CCA gene-based test nonetheless provides a computationally efficient alternative when analyzing such traits. In the multivariate scenario, a significant result can be followed by (i) an examination of the weights attributed by the CCA to each phenotype; and/or (ii) inspection of the univariate gene-based results. In this way, the individual traits underlying the multivariate association can be identified.

As proof-of-principle, we applied the CCA gene-based test to the analysis of six WBC traits measured in up to 1061 unrelated healthy individuals. Univariate gene-based analyses of both common and uncommon SNPs identified two notable regions of association, one for lymphocyte counts on chromosome 19 that included the two paralog genes SAFB and SAFB2, and another for neutrophil counts including several genes in the histone 1 cluster on chromosome 6. SAFB is involved in the transcription repression of immune regulatory genes, including HLA and cytokine genes (Hammerich-Hille et al., 2010). Furthermore, SAFB$^{-/-}$ knockout mice show defects in the development of the hematopoietic system, increased WBC counts, hypoplasia of the thymus and increased signs of infections (Greene and Oestreicher, 2010). Thus, together with these data, our results suggest that SAFB and/or SAFB2 represent novel putative genes with multiple independent QTL regulating variation in lymphocyte counts in humans. Our analyses identified two specific uncommon variants (rs4239608 and rs806706) in this region that could be the focus of subsequent follow-up studies.

The second notable region was the histone 1 gene cluster on chromosome 6, which was associated with neutrophil counts. Histones are increasingly recognized to have an important role in cellular signaling and innate immunity, in addition to their role in the regulation of chromatin structure (Parseghian and Luhrs, 2006). Interestingly, mice challenged with a sublethal dose of histones, develop a prominent accumulation of neutrophils in the lung (Xu et al., 2009). Furthermore, the histone 1 gene cluster contains the greatest enrichment of SAFB target-binding sites (Hammerich-Hille et al., 2010). These results thus raise the possibility that genetic variants in this region may affect the binding of SAFB which, in turn, may regulate the transcription of key histone genes involved in neutrophil production, migration or apoptosis. Candidate SNPs that may explain this effect include rs1458378, rs17598658 and rs4593350; further studies that attempt to replicate our observed association between these and neutrophil counts are warranted.

In conclusion, we show that CCA provides a useful framework for a gene-based test of association which, in some situations, may outperform commonly used permutation-based approaches. Our
C.S.Tang and M.A.R. Ferreira

analysis of six hematology traits identified novel putative genes harboring multiple independent QTL regulating lymphocyte and neutrophil counts, but these need to be confirmed by follow-up studies.

ACKNOWLEDGEMENTS

We thank Shaun M. Purcell for some of the original ideas we developed in this work. We also thank Sarah E. Medland, Scott Gordon, Dizie Statham, Ann Eldridge, Marlene Grace, Lisa Bowdler, Steven Crooks, Harry Breeby, Anjali K. Henders, Dale R. Nyholt, Peter M. Visscher, Ian H. Frazer, Grant W. Montgomery and Nicholas G. Martin, who were part of the original team that ascertained, phenotyped, genotyped and processed SNP data for the Brisbane twins and their families who participated in the hematology study.

Funding: Australian National Health and Medical Research Council (grants 241944, 339462, 389875, 389891, 389938, 443036, 442915, 442981, 496739, 552485, 552498, 613679), Australian Research Council (A7960034, A79906588, A79801419, DP0212016, DP0343921).

Conflict of Interest: none declared.

REFERENCES


