BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 15 2012, pages 2084–2085
doi:10.1093/bioinformatics/bts315

Genetics and population analysis

Advance Access publication May 26, 2012

Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies

Robert Kofler and Christian Schlötterer*

Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: An analysis of gene set [e.g. Gene Ontology (GO)] enrichment assumes that all genes are sampled independently from each other with the same probability. These assumptions are violated in genome-wide association (GWA) studies since (i) longer genes typically have more single-nucleotide polymorphisms resulting in a higher probability of being sampled and (ii) overlapping genes are sampled in clusters. Herein, we introduce Gowinda, a software specifically designed to test for enrichment of gene sets in GWA studies. We show that GO tests on GWA data could result in a substantial number of false-positive GO terms. Permutation tests implemented in Gowinda eliminate these biases, but maintain sufficient power to detect enrichment of GO terms. Since sufficient resolution for large datasets requires millions of permutations, we use multi-threading to keep computation times reasonable.

Availability and implementation: Gowinda is implemented in Java (v1.6) and freely available on http://code.google.com/p/gowinda/

Contact: christian.schloetterer@vetmeduni.ac.at

Supplementary information: Manual: http://code.google.com/p/gowinda/wiki/Manual. Test data and tutorial: http://code.google.com/p/gowinda/wiki/Tutorial. Validation: http://code.google.com/p/gowinda/wiki/Validation.

Received on March 12, 2012; revised on May 3, 2012; accepted on May 22, 2012

1 INTRODUCTION

The advent of high-throughput analysis such as single-nucleotide polymorphism (SNP) arrays and next-generation sequencing enabled large-scale genome-wide association (GWA) studies (Nordborg and Weigel, 2008) or GWA-like studies, such as selective genotyping (Darvasi and Soller, 1994) and experimental evolution (Turner *et al*., 2011), for almost any phenotype of interest. These studies typically yield hundreds of candidate SNPs associated with the studied trait. A wide-spread approach to shed light on the biological implications of these SNPs is a test for gene set enrichment [e.g. Gene Ontology (GO)] (Ashburner *et al*., 2000; Wang *et al*., 2010).

Such an analysis of gene set enrichment is based on the assumptions that all genes are sampled independently from each other with the same probability. These assumptions are violated with data from GWA studies as (i) longer genes usually have more SNPs resulting in a higher probability of being sampled and (ii) overlapping genes are sampled in clusters (Holmans *et al*., 2009). For these reasons, we developed Gowinda, a user-friendly and multi-threaded software specifically designed for detecting unbiased enrichment in gene sets from large datasets such as generated by GWA studies. By relying on standard file formats, any species with a sequenced and annotated genome may be analyzed, hence it favorably compares to a similar tool that is restricted to the analysis of GWA datasets in humans (Holmans *et al*., 2009). [For a review of different available methods, see Wang *et al*. (2010)]. We validated Gowinda and show that the biases inherent to GWA dataset could result in a substantial number of false-positive GO terms and that Gowinda eliminates these biases while still yielding highly reliable results.

*To whom correspondence should be addressed.

2 IMPLEMENTATION

Gowinda calculates the significance of overrepresentation for each gene set with permutation tests. Gowinda randomly samples SNPs from the total set of SNPs and records the associated genes. After repeating this permutation multiple times, an empirical null distribution of gene abundance for every gene set is obtained. The significance of overrepresentation of the candidate SNPs is estimated from the empirical null distribution. To account for multiple testing, an empirical false discovery rate (FDR) is calculated, by dividing the number of expected gene sets for a given *P*-value (averaged from the simulations) by the number of observed gene sets.

Gowinda requires four input files, all of which widely used standard formats: a file with the annotation of the species of interest (.gtf or .gff), a file with the total set of SNPs used for the GWA study [.vcf, .mpileup or similar (Danecek, *et al*., 2011; Li, *et al*., 2009)], a file containing the candidate SNPs (must be a subset of all SNPs) and a file with the mapping of genes to gene sets. Such files with the mapping between genes and GO terms can, for example, be obtained either from FuncAssociate2 (Berriz *et al*., 2009) or from High-Troughput (HT) GoMiner (Zeeberg *et al*., 2005). In addition to this, Gowinda can be used to identify enrichment of SNPs in any user-defined gene set (Manual: http://code.google.com/p/gowinda/wiki/Manual; Test data and walkthrough: http://code.google.com/p/gowinda/wiki/Tutorial).

Gowinda does not reproduce the exact pattern of linkage disequilibrium (LD) between SNPs but offers two complementary test strategies making two extreme assumptions about LD:

- SNPs are in linkage equilibrium: Gowinda randomly samples the same number of SNPs as candidate SNPs. Subsequently, the corresponding genes are identified and overrepresentation is estimated as described above. Note that the number of randomly sampled SNPs is kept constant in the simulations

(*–mode SNP*) and that a gene may be considered several times according to the number of candidate SNPs.

- SNPs are in complete LD: Gowinda randomly samples SNPs until the corresponding number of genes is identical to the number candidate genes. Finally, the significance of overrepresentation is estimated as described above. Note that the number of randomly sampled genes is kept constant in the simulations (*–mode gene*), but the number of sampled SNPs may vary between simulations. Furthermore, every gene is only considered once even when containing several candidate SNPs. This approach assumes complete LD between SNPs within a gene but does, however, not account for LD between genes.

Another important feature of Gowinda is a flexible definition of a gene. This enables the user to include SNPs mapping to different features in the analysis, such as exons, introns, untranslated region or 2000 bp downstream. Possible definitions are as follows: exons, CDS, exons + introns, untranslated region, upstream + exons + introns, exons + introns + downstream and upstream + exons + introns + downstream.

3 VALIDATION AND BENCHMARKS

To test the reliability of Gowinda, we asked whether Gowinda reproduces the significance of overrepresentation for GO categories as the widely used tool HT GoMiner (Zeeberg *et al*., 2005). We created an unbiased dataset by filtering for non-overlapping *Drosophila melanogaster* genes and introduced exactly five SNPs into each of the genes. Subsequently, we randomly sampled 1000 SNPs and computed the significance for the overrepresentation of every GO category, either on the basis of SNPs using Gowinda or based on the corresponding genes using HT GoMiner. We found that Gowinda yields almost identical results as HT GoMiner (Fig. 1A; Spearman's rank correlation; $\rho > 0.99$; P-value $< 2.2 \times -10^{-16}$). We also assessed the bias introduced in GWAS data and to what extent Gowinda corrects for this bias. We created a biased dataset by introducing a SNP every 100 bp into all genes and again randomly sampled 1000 SNPs. We calculated the significance of overrepresentation using Gowinda and HT GoMiner. Consistent with the expected bias due to different gene lengths and overlapping genes, HT GoMiner reports a significant enrichment for 341 GO categories (FDR < 0.05), whereas Gowinda correctly reports zero (FDR < 0.05). We also found that the correlation of the P-values of overrepresentation for Gowinda and HT Gominer dramatically decreased with the biased dataset (Fig. 1B; Spearman's rank correlation; $\rho > 0.56$; P-value $< 2.2 \times 10^{-16}$).

Finally, we tested whether Gowinda correctly identifies overrepresentation for five randomly preselected GO categories. We randomly picked five small GO categories (5–10 genes) and introduced a candidate SNP into every gene associated with these GO categories. Subsequently, we randomly sampled SNPs from the biased dataset until a total of 1000 candidate SNPs were obtained.

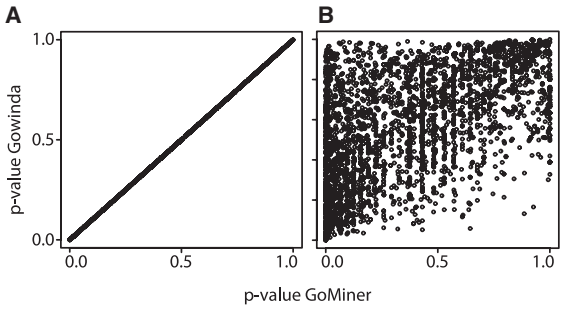

Fig. 1. Correlation of the significance of overrepresentation for GO terms as calculated by Gowinda and GoMiner using an unbiased (**A**) and a biased (**B**) dataset

After analysis for GO term enrichment, we found that Gowinda correctly identified all preselected GO categories (FDR < 0.05). Interestingly, significant enrichment was also identified for another 14 GO categories, which is due to the nesting of GO categories. Details about validation can be found at: http://code.google.com/p/gowinda/wiki/Validation.

Gowinda is reasonably fast, in *D. melanogaster* 1 000 000 simulations for 2 000 candidate SNPs out of a total of 1.8 million SNPs take about 31 min with a Mac Pro (10.5.8) using eight threads and requires about 1.2 GB of RAM. Memory consumption is mostly dependent on the total number of SNPs and computation time scales with the number of simulations.

Funding: Austrian Science Fund (FWF) grant (P19467) to C.S.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Berriz,G.F. *et al*. (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043–3044.

Danecek,P. *et al*. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Darvasi,A. and Soller,M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*, **138**, 1365–1373.

Holmans,P. *et al*. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Nordborg,M. and Weigel,D. (2008) Next-generation genetics in plants. *Nature*, **456**, 720–723.

Turner,T.L. *et al*. (2011) Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.*, **7**, e1001336.

Wang,K. *et al*. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 843–854.

Zeeberg,B.R. *et al*. (2005) High-throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics.*, **6**, 168.