Bioinformatics, 35(1), 2019, 160–163
doi: 10.1093/bioinformatics/bty529
Advance Access Publication Date: 13 July 2018
Applications Note

OXFORD

Genetics and population analysis

GIGSEA: genotype imputed gene set enrichment analysis using GWAS summary level data

Shijia Zhu[1,*], Tongqi Qian[1], Yujin Hoshida[2], Yuan Shen[3], Jing Yu[4] and Ke Hao[1,5,*]

[1]Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA and [2]Liver Tumor Translational Research Program, Simmons Comprehensive Cancer Center, Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA and [3]Department of Psychiatry and [4]Department of Ophthalmology and [5]Department of Respiratory Medicine, Shanghai Tenth People's Hospital, Tongji University, Shanghai, 200092, China

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on November 21, 2017; revised on May 11, 2018; editorial decision on June 26, 2018; accepted on July 11, 2018

Abstract

Summary: Summary level data of GWAS becomes increasingly important in post-GWAS data mining. Here, we present GIGSEA (Genotype Imputed Gene Set Enrichment Analysis), a novel method that uses GWAS summary statistics and eQTL to infer differential gene expression and interrogate gene set enrichment for the trait-associated SNPs. By incorporating empirical eQTL of the disease relevant tissue, GIGSEA naturally accounts for factors such as gene size, gene boundary, SNP distal regulation and multiple-marker regulation. The weighted linear regression model was used to perform the enrichment test, properly adjusting for imputation accuracy, model incompleteness and redundancy in different gene sets. The significance level of enrichment is assessed by the permutation test, where matrix operation was employed to dramatically improve computation speed. GIGSEA has appropriate type I error rates, and discovers the plausible biological findings on the real data set.
Availability and implementation: GIGSEA is implemented in R, and freely available at www.github.com/zhushijia/GIGSEA.
Contact: shijia.zhu@mssm.edu or ke.hao@mssm.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWAS) have been successfully applied to many diseases and traits. However, the genome-wide significant variants only account for a minority of the trait heritability, whereas SNPs below the genome-wide significance level may also harbor real association. Enrichment analysis is an important tool in prioritizing variants from GWAS. Instead of conducting analysis for every single SNP or gene, enrichment analysis tests disease association at the level of a group of functionally related SNPs or genes, such as those belonging to the same biological pathway. These approaches aim to increase power by combining association signals. Further, gene set analysis can also shed more light on the biological processes underlying complex diseases. Many methods have been proposed (de Leeuw *et al.*, 2015; Lee *et al.*, 2012), however, many challenges and limitations remained, for example, arbitrary thresholds in selecting GWAS loci, differentially expressed genes, gene boundaries, long-range regulation and multiple-marker regulation for effective snps, and how to adjust various bias from the gene size, linkage disequilibrium (LD) and the redundancy among gene sets.

The expression quantitative trait locus (eQTL) of the disease relevant gene provides an empirical and unbiased link between GWAS loci and genes potentially mediating the genetic effects (Nicolae *et al.*, 2010). Recently, several methods, e.g. TWAS (Gusev *et al.*, 2016), PrediXcan (Gamazon *et al.*, 2015) and MetaXcan (Barbeira *et al.*, 2016) were proposed to integrate GWAS summary data with eQTL to impute transcription levels of disease-associated genes, providing a path to aggregate multiple markers at the gene level. However, genes identified by these methods are still picked with arbitrary thresholds and then enter the enrichment analysis. Herein, we proposed a novel method for the gene set enrichment analysis on GWAS summary data, called GIGSEA (Fig. 1), which is an elegant extension to TWAS and MetaXcan. GIGSEA carried out the enrichment analysis on the imputed gene expression, where the inheritability and imputation uncertainty were accounted for by the weighted linear regression model.

2 Methods

2.1 Genotype imputed differential gene expression

We used MetaXcan (Barbeira *et al.*, 2016) to impute the trait-associated differential gene expression (DGE). MetaXcan integrates GWAS summary with eQTL to map trait-associated genes. The eQTL dataset or the 1000 Genomes can be used as LD reference among markers. The eQTL summary was pre-calculated from gene expression studies, such as GTEx (Lonsdale *et al.*, 2013) and Depression Genes and Networks (DGN, blood tissue eQTL) (Battle *et al.*, 2014). Users provide the GWAS summary data to impute the genetically regulated gene expression and conduct gene set enrichment test.

2.2 Enrichment analysis

Based on the imputed DGE (Z-score), we used the linear regression model to build a threshold-free gene set enrichment test, examining whether genes are overrepresented in a particular gene set. Apparently, the gene expression cannot be perfectly predicted only using genotype, and the uncertainty is quantified as correlation (r^2) between the measured and predicted gene expression from cross-validation by MetaXcan. We took the uncertainty as weights, building a weighted linear regression model. Two kinds of tests were developed (detail in Supplementary Material):

1. Single gene-set enrichment analysis (SGSEA)

For each gene-set, we built a weighted simple linear regression model, regressing the imputed gene expression on that gene set.

2. Multiple gene-set enrichment analysis (MGSEA)

To address the redundancy among gene sets, we built a weighted multiple linear regression model, taking into account all gene sets in one model. The redundancy in one gene set can be adjusted by considering all other gene sets as covariates.

2.3 Permutation test to assess significance level

The weights used in the regression model make the regression *residual* potentially deviate from the homoscedasticity assumption, resulting in a non-uniform distribution of *P*-values under the null (Supplementary Material). Therefore, we used the permutation test to assess the *P*-values of regression coefficients from the weighted regression model. We repeatedly randomized the imputed DGE to obtain a global null distribution of no associated gene sets and calculated the empirical *P*-value for each gene set. For the GIGSEA, especially with many gene sets tested, a large number of regression models would be interrogated, resulting in very intensive computation. To speed up, the large matrix operation was used in GIGSEA. For SGSEA, we used the weighted Pearson correlation to rank the regression coefficient, as they take the same test statistic, and furthermore, we expressed the weighted Pearson correlation in terms of large matrix inner product to calculate all correlations in one step, therefore substantially improving the time efficiency. Taking GO (16 339 GO terms) as an example, the weighted single regression model takes ~3.6 days to run 1000 permutations (a single Intel i7 2.1 G CPU and 16 G memory), while the weighted Pearson correlation takes only 2.5 min, improving the efficiency by ~2000 times. Likewise, to accelerate MGSEA, we used the matrix solution of the weighted multiple linear regression model, which also largely improves the time efficiency. Furthermore, to correct the multiple hypothesis testing, we used the Bayes Factor, which accounts for both local fdr (Efron and Tibshirani, 2002) and prior odds ratio between null and nonnull classes (details in Supplementary Material).

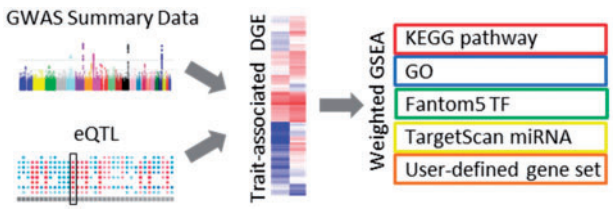

Fig. 1. The flowchart of GIGSEA. GIGSEA performs two levels of aggregation: from SNPs (GWAS and eQTL) to genes, and from genes to gene sets

2.4 Gene sets

GIGSEA used weighted linear regression model to perform gene set enrichment test, so, it allowed both discrete and continuous-valued gene sets. Multiple collections of gene sets have already been incorporated in the current tool:

1. Discrete-valued gene set: MsigDB (Subramanian *et al.*, 2005) (186 KEGG pathways, binding targets of 221 miRNAs and 615 transcriptional factors), and 16 339 Gene Ontology terms (GO, addressed the offspring gene sets);
2. Continuous-valued gene set: Fantom5 promoter based binding target prediction for 500 PWMs of transcriptional factors (Pachkov *et al.*, 2013), and TargetScan binding target prediction for 87 miRNA seeds (Friedman *et al.*, 2009);
3. Users can also provide their own gene sets of interest.

3 Simulation study on type I error rate

In order to assess the type I error rates of GIGSEA, we simulated a dataset, for which no enriched gene sets are expected, from the real psychiatry disease GWAS (Consortium, 2014) as the following. The psychiatry GWAS summary data and the pre-calculated DGN eQTL database were used as input for MetaXcan to impute the psychiatry-associated DGE, where the mean MetaXcan prediction r^2 is 12.5%, 60.4% SNPs were used in the model, and 11 230 genes were predicted. Next, we randomized the genes in the DGE but keep the gene-weight pairs intact. We took MSigDB KEGG pathway as an example, and applied the GIGSEA to the shuffled DGE, to investigate the false positive enriched pathways by setting the alpha level as 5%. We repeated the shuffling for 10 times, and treated the average false positive rate as type I error rate. Based on the simulation, we evaluated the type I error rates of four analyses: SGSEA without

weights, MGSEA without weights, SGSEA with weights and MGSEA with weights. For each analysis, we calculated the empirical *P*-values based on 10 000 times permutation. We found that all type I error rates were below the alpha level.

4 Application to a cardiovascular GWAS dataset

We applied GIGSEA to cardiovascular disease (CVD) GWAS, CARDIoGRAMplusC4D (60 801 cases, 123 504 controls and 9.4 M SNPs) (Nikpay *et al.*, 2015). To impute the genes associated with CVD, we used DGN blood eQTL (Battle *et al.*, 2014) and the 1000 Genomes as LD reference. 11 537 genes were imputed with high quality prediction. On average, the gene expression prediction r^2 is 12.4%, and the majority of SNPs (98.88%) used for prediction are available in the CARDIoGRAMplusC4D GWAS. We tested different classes of gene sets (Supplementary Tables and Supplementary Material) and yield biologically plausible findings with convincing literature support:

- MSiGDB KEGG pathway: out of 186 pathways, SGSEA detected 6 significantly enriched pathways (empirical *P*-value < 0.05 and BayesFactor > 3), e.g. KEGG vascular smooth muscle contraction, and all of them are supported by literatures, while MGSEA found two significantly enriched pathways (both overlap with SGSEA), and both of them are supported by biochemical knowledge and literatures.
- Fantom5 transcriptional factor: out of 500 transcriptional factors, SGSEA found six significant TFs, out of which four are supported by literatures, while MGSEA found six significant TFs (two overlap with SGSEA), out of which three are supported;
- TargetScan miRNA: out of 87 miRNAs, SGSEA found six significant enriched miRNAs and five are supported by literatures, e.g. miR-138 and miR-216, while MGSEA found five significant miRNAs (three overlap with SGSEA), and four are supported.
- GO: the number of GO terms is close to or even larger than the imputed genes, making the regression model difficult to be estimated, and therefore, we only performed SGSEA. The top ranking enriched GO terms are GO: 0031116 (Positive regulation of microtubule polymerization), GO: 0001818 (Negative regulation of cytokine production), and similar GO terms. However, they fail to survive the heavy multiple testing correction, due to the large number of GO terms. In experimental studies, microtubules were shown to be accumulated, thereby impeding sarcomere motion and promoting cardiac dysfunction, and the pro-inflammatory cytokines are also found to be involved in cardiac depression and in the complex syndrome of heart failure.

Although the empirical *P*-values by SGSEA and MGSEA demonstrated largely consistent trend, they also revealed different gene sets (Supplementary Fig. S1a–c), e.g. KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450 (SGSEA empirical *P*-value = 4.4e-03, MGSEA empirical *P*-value = 0.35). Such different significance is mainly because of the sharing information between KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450 and multiple other pathways (Pearson correlation $r > 0.22$ with *P*-value $<$ 2.2e-16, Supplementary Fig. 1d), which are also associated with the disease, and furthermore, different from the SGSEA, the MGSEA can regress away the sharing information, with the *P*-values calculated only for the unique information. In addition, the MGSEA also found significant gene sets which failed to be detected by the SGSEA, including the FATTY_ACID_METABOLISM pathway (MGSEA empirical *P*-value = 1.0e-02), the ESRRA_ESR2 TF (MGSEA empirical *P*-value = 5.0e-04), and the hsa-miR-128 miRNA (MGSEA empirical *P*-value = 6.4e-03). Their involvements in the cardiac disease are also supported by literatures. This fact suggests that the MGSEA can remove the redundant information existing in gene sets, thereby enabling a better detection of the true associations. This is quite similar to the adjustment for covariates in the GWAS study. Taken together, we showed the effectiveness of both SGSEA and MGSEA and also their complementary capabilities in uncovering the trait or disease relevant gene sets.

5 Comparison with another GWAS-based GSEA tool: FUMA

There are few GWAS-based gene set enrichment methods, which use both GWAS summary data and eQTL information. We only compared to a very recent published online tool: Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) (Watanabe *et al.*, 2017), which can map SNP-gene pairs based on both position [annotations from ANNOVAR (Wang *et al.*, 2010)], and eQTL (independent significant SNPs in user-defined tissues). We analyzed the CVD GWAS using FUMA. To make fair comparison, we employed the blood eQTL in FUMA (same tissue type as in GIGSEA analysis), and investigated the enriched gene sets for MSigDB KEGG pathways. Compared to GIGSEA, FUMA found different and less KEGG pathways: glycerolipid metabolism (adj*P*-value = 0.040), beta alanine metabolism (adj*P*-vlaue = 0.029), and selenoamino acid metabolism (adj*P*-value = 0.041). Importantly, we tested FUMA with and without eQTL mapping (without eQTL, FUMA solely relies on positional mapping), and surprisingly, FUMA returned exactly the same pathways, suggesting that the eQTL information does not significantly contribute to the enrichment analysis in FUMA. Further, GIGSEA incorporated more comprehensive information into a regression model for enrichment test than FUMA, including both association strength and inheritabilities.

6 Summary

GIGSEA uses GWAS summary data and disease relevant eQTL for gene set enrichment analysis. GIGSEA addressed such challenges in SNP enrichment as gene size, gene boundary, SNP long-range regulation, multiple-marker regulation, arbitrary cutoff on GWAS and eQTL effect size or significance. Our method can be viewed as an extension of the SNP-imputed gene-level test (e.g. TWAS and PrediXcan) into the gene set level. It is a timely tool leveraging three recent advances in genetic study (i) availability of summary level data of large GWAS of many diseases/traits; (ii) availability of eQTL data of disease relevant tissues, e.g. GTEx (Lonsdale *et al.*, 2013) and (iii) new approaches to impute the gene expression level (e.g. MetaXcan). Application to real data demonstrated the good performance and discovered the biologically meaningful findings. GIGSEA is based on permutation test, and extensively optimized for time efficiency. GIGSEA would have wide utility in the post-GWAS era to mine GWAS summary data and reveal molecular mechanisms mediating genetic predisposition of diseases.

Funding

Partially supported by National Natural Science Foundation of China (No. 21477087, 91643201), Minister of Science and Technology of China (2016YFC0206507), NIH/NIDDK (R01DK106593 and U24DK062429), and NIH/NIEHS (1R01ES029212-01).

Conflict of Interest: none declared.

References

Barbeira,A. *et al*. (2016) Integrating tissue specific mechanisms into GWAS summary results. *bioRxiv*, **045260**.

Battle,A. *et al*. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*., **24**, 14–24.

Consortium,SWGotPG. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

de Leeuw,C.A. *et al*. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol*., **11**, e1004219.

Efron,B. and Tibshirani,R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol*., **23**, 70–86.

Friedman,R.C. *et al*. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*., **19**, 92–105.

Gamazon,E.R. *et al*. (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet*., **47**, 1091–1098.

Gusev,A. *et al*. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet*., **48**, 245–252.

Lee,P.H. *et al*. (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, **28**, 1797–1799.

Lonsdale,J. *et al*. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet*., **45**, 580–585.

Nicolae,D.L. *et al*. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*., **6**, e1000888.

Nikpay,M. *et al*. (2015) A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet*., **47**, 1121.

Pachkov,M. *et al*. (2013) SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*., **41**, D214–D220.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., **38**, e164–e164.

Watanabe,K. *et al*. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun*., **8**, 1826.