Bioinformatics, 34(2), 2018, 286–288
doi: 10.1093/bioinformatics/btx509
Advance Access Publication Date: 14 September 2017
Applications Note

OXFORD

Genome analysis

JEPEGMIX2: improved gene-level joint analysis of eQTLs in cosmopolitan cohorts

Chris Chatzinakos[1,*], Donghyung Lee[2], Bradley T. Webb[1], Vladimir I. Vladimirov[1], Kenneth S. Kendler[1] and Silviu-Alin Bacanu[1]

[1]Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298, USA and [2]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 13, 2017; revised on July 10, 2017; editorial decision on August 7, 2017; accepted on September 13, 2017

Abstract

Motivation: To increase detection power, researchers use gene level analysis methods to aggregate weak marker signals. Due to gene expression controlling biological processes, researchers proposed aggregating signals for expression Quantitative Trait Loci (eQTL). Most gene-level eQTL methods make statistical inferences based on (i) summary statistics from genome-wide association studies (GWAS) and (ii) linkage disequilibrium patterns from a relevant reference panel. While most such tools assume homogeneous cohorts, our **G**ene-level **J**oint **A**nalysis of functional SNPs in **C**osmopolitan **C**ohorts (JEPEGMIX) method accommodates cosmopolitan cohorts by using heterogeneous panels. However, JEPGMIX relies on brain eQTLs from older gene expression studies and does not adjust for background enrichment in GWAS signals.

Results: We propose JEPEGMIX2, an extension of JEPEGMIX. When compared to JPEGMIX, it uses (i) cis-eQTL SNPs from the latest expression studies and (ii) brains specific (sub)tissues and tissues other than brain. JEPEGMIX2 also (i) avoids accumulating averagely enriched polygenic information by adjusting for background enrichment and (ii) to avoid an increase in false positive rates for studies with numerous highly enriched (above the background) genes, it outputs gene *q*-values based on Holm adjustment of *P*-values.

Availability and implementation: https://github.com/Chatzinakos/JEPEGMIX2.

Contact: chris.chatzinakos@vcuhealth.org

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

1 Introduction

Gene expression is believed to have influenced human evolution and play a key role in diseases (Emilsson *et al.*, 2008). Thus, it is critical for understanding diseases and developing treatments. The importance of gene expression was further underlined by the enrichment of association signals in SNPs tagging gene expression (Nica and Dermitzakis, 2008; Nicolae *et al.*, 2010), which are denoted as expression quantitative trait loci (eQTL).

Currently, the identification of complex disease susceptibility loci is performed via genome-wide association studies (GWAS). It involves scanning single nucleotide polymorphisms (SNPs) across the entire genome for genetic variants associated with a trait. Univariate analysis of GWAS is still the de facto tool for identifying trait associated SNPs (Wellcome Trust Case Control, 2007). However, when analyzing more complex GWAS SNPs with weak or moderate effect sizes, the significant findings account only for a small fraction of the total trait variation (Manolio *et al.*, 2009). Due to their small effect sizes, these SNPs are rarely detected in GWAS (Yang *et al.*, 2010). To increase the power of detection, researchers proposed analyzing genetic variants multivariately (Wang *et al.*, 2007).

One type of multivariate analyses is the transcriptome-wide association study (TWAS) which identifies significant expression-trait associations. Such methods, e.g. joint effect on phenotype of eQTL/

functional SNPs associated with a gene (JEPEG) (Lee *et al.*, 2015), PredictXcan (Gamazon *et al.*, 2015), JEPEGMIX (Lee *et al.*, 2016) and TWAS (Gusev *et al.*, 2016) use eQTL to predict gene expression and/or infer which genes are associated with traits. However, unlike competing non-eQTL paradigms, e.g. LDscore/LDpred (Bulik-Sullivan *et al.*, 2015), current TWAS methods (i) lack competitive adjustment for background enrichment ('average signal') and (ii) do not output q-values that control false positive rates when there is a substantial number of genes enriched (above background) in signals.

To address these shortcomings, we propose JEPEGMIX2, an extension of JEPEGMIX, which, in addition to the existing advantage of imputing eQTLs statistics and inferring gene-trait association in cosmopolitan cohorts, it also (i) adjusts for background enrichment, (ii) offers the option to upweight rarer eQTLs and (iii) to avoid false positive rate increase for high signal enrichment, it outputs Holm q-values.

2 Materials and methods

To avoid a mere accumulation of just averagely enriched polygenic information, we competitively adjust χ^2 statistics for background enrichment. This is achieved by adjusting the statistic for average non-centrality. Such 'centralized' JEPEGMIX statistic we denote as competitive (C) and the original statistic as the non-competitive (NC).

Let Z be the vector of Z-scores for measured SNPs in the genome scans. Due to polygenicity, the expected genome scan $\chi^2_1 = Z^2$ statistics, each with 1 degree of freedom (df), has a non-zero background noncentrality parameter λ^2, i.e. $E(Z^2) = 1 + \lambda^2$. Thus, by the method of moments, we can estimate $\widehat{\lambda^2} = \bar{Z}^2 - 1$, where $\bar{Z}^2$ is computed using all measured SNPs in the genome scan, However, given that $\lambda^2 \geq 0$, a better estimator is, thus, $\widehat{\lambda^2} = \max(\bar{Z}^2 - 1,\ 0)$. To develop a competitive test, before computing gene-level statistics, Z-scores must be shrunk towards zero by adjusting for the average background enrichment. This can be achieved via a 3 step process:

1. Recompute, under 'average' noncentrality, the P-value associated with χ^2_1 statistics: $P' = 1 - F(Z^2|\ \widehat{\lambda^2})$, where $F(.|\ \widehat{\lambda^2})$, is the cumulative distribution function (cdf) of the non-central χ^2_1 distribution with 1 df and noncentrality parameter $\widehat{\lambda^2}$.
2. Transform P' into its quantile vector from a central χ^2_1 distribution with 1 df, i.e. $\chi^2 = F^{-1}(1 - P'|\ \lambda^2 = 0)$,
3. Transform χ^2 to a 'central' Z-score: $Z' = sign(Z) * \sqrt{\chi^2}$.

By Delta method (a first order Taylor approximation), Z' as a linear transformation (deflation) of Z has the same correlation structure. Thus, Z' can be used to build the competitive gene statistics (Supplementary Text S1), which has the same variance as their non-competitive versions.

To facilitate user-specific input along with future extensions, the new annotation file now includes a R-like formula for the expression of each gene as a function of its eQTL genotypes. The annotation file includes cis-eQTL for all tissues available in PREDICTDB (http://predictdb.hakyimlab.org/). To avoid making inference about genes poorly predicted by SNPs, for the 44 available tissues we retain only genes for which the expression is predicted with q-value < 0.05 from its eQTLs. Additionally, given the increased deleteriousness of rarer mutations, we offer the possibility to upweight coefficient of rarer variants (Supplementary Text S1 for statistic computation) using a Madsen and Browning type approach (Madsen and Browning, 2009). For linkage disequilibrium (LD) estimates in cosmopolitan cohorts (needed for both imputation and statistical inference), we allow user to input the study cohort proportions of ethnicities from the reference panel. LD patterns of the study cohort are estimated as a weighted mixture (with the above weights) of the LD matrices for all ethnic groups in a reference panel (Supplementary Text S2). LD patterns are subsequently used to (i) accurately impute summary statistics of unmeasured eQTLs (Supplementary Text S3) and (ii) compute the variance of the SNP linear combinations used for gene level tests in each tissue (Supplementary Text S2). The current version uses the 1000 genome (1KG) Phase I release version 3 as reference panel (Durbin *et al.*, 2010). It consists of 379 Europeans, 286 Asians, 246 Africans and 181 Native Americans.

3 Simulations

To estimate the false positive rates of JEPEGMIX2, for five different cosmopolitan studies scenarios (Supplementary Text S4), we simulated (under H_0) 100 cosmopolitan cohorts of 10, 000 subjects for Ilumina 1 M autosomal SNPs using 1KG haplotype patterns (Supplementary Text S4, Supplementary Table 1). The subject phenotypes were simulated independent of genotypes as a random Gaussian sample. SNP phenotype-genotype association summary statistics, were computed as a correlation test. We obtained JEPEGMIX2 statistics for: (i) competitive (C), non-competitive (NC) and (ii) tests with rare (Madsen and Browning like) (R) and non-rare (NR) eQTL weights. To test the ability of methods to maintain false positive rates under background enrichment, we provide an enriched scenario. Under this scenario, we quantile transform the simulated 'central' Z-score (CZ) to a 'non-central' Z-score (NCZ) scenario by following the three steps from the previous section with the first step having noncentrality $\lambda^2 = 0$ and the second one $\lambda^2 = 0.5$ [extrapolation of PGC3 Schizophrenia nocentrality from PGC2 λ^2]{\booklink="DPDFMK55"} (Ripke *et al.*, 2013)]. We also applied JEPEGMIX2 to 16 real summary datasets (Supplementary Text S5, Supplementary Table S2). To limit the increase in Type I error rates of JEPEGMIX2, we deem as significantly associated only genes with Holm-adjusted P-value (q-value) < 0.05. Due to C4 explaining most of Major Histocompatibility (MHC) (chr6: 25–33 Mb) (McCarthy *et al.*, 2016), signals for schizophrenia (SCZ), for this trait, we omit non-C4 genes in this region.

4 Results

JEPEGMIX2 with competitive (C) statistics, controls the false positive rates at or below nominal thresholds for both central (CZ) and non-central (NCZ) scenarios while the non-competitive (NC) has similar behavior only for the central case (when the GWAS statistics are not enriched) (Supplementary Text S5, Supplementary Figs S1–S5). Under the enriched scenario (NCZ) the non-competitive version of the test has much increased false positive rates.

Table 1. Signals for real datasets

Traits	No unique genes
SCZ	68
ALZ	34
AMD	17
BIP	11
HDL	79
LDL	78
T2D	5
TG	48
Smoking	5

Using the Holm *P*-value adjustment and both rare (R) and non-rare (NR) e QTL weights, for the real datasets significant gene signals were found in 9 traits, for which we present heatmaps (Supplementary Text S5, Supplementary Figs S6–S23). The number of genes with q-value < 0.05 is presented in Table 1 (for the abbreviations see Supplementary Table S2). Each analysis ran in less than 3 h on a cluster node with 4× Intel Xeon 6 core 2.67 GHz.

5 Conclusions

We propose JEPEGMIX2, an updated software/method for testing the association between (cis-eQTL mediated) gene expression and trait. Unlike existing methods, even for highly enriched GWAS, JEPEGMIX2 competitive version fully controls the false positive rates at or below nominal levels. To the applicability of JEPEGMIX to cosmopolitan cohorts, we add a competitive version and extend the number of included (i) eQTLs and (ii) tissues. Unlike existing methods, it also accommodates up weighting of the rare variants and avoids the increased rate of false positives incurred by FDR adjustment (under enrichment) by using a Holm adjustment. While gene expression in different tissues are often correlated and incomplete due to the rather small sample sizes of existing gene expression experiments, the capacity of discriminating causal tissues will be enhanced by further increases in sample size of such studies. Being written in C++, JEPEGMIX2 is very fast. Future versions of the software will use larger reference panels.

Conflict of Interest: none declared.

References

Bulik-Sullivan,B.K. *et al*. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet*., **47**, 291–295.

Durbin,R.M. *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Emilsson,V. *et al*. (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.

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) Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nat. Commun*., **7**, 10979.

Lee,D. *et al*. (2015) JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics*, **31**, 1176–1182.

Lee,D. *et al*. (2016) JEPEGMIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts. *Bioinformatics*, **32**, 295–297.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*., **5**, e1000384.

Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

McCarthy,S. *et al*. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*., **48**, 1279–1283.

Nica,A.C. and Dermitzakis,E.T. (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet*., **17**, R129–R134.

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.

Ripke,S. *et al*. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet*., **45**, 1150–1159.

Wang,K. *et al*. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet*., **81**, 1278–1283.

Wellcome Trust Case Control (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Yang,J. *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet*., **42**, 565–569.