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

Chris Chatzinakos

Donghyung Lee
*The Jackson Laboratory, donghyung.lee@jax.org*

Bradley T Webb

Vladimir I Vladimirov

Kenneth S Kendler

*See next page for additional authors*

Follow this and additional works at: [https://mouseion.jax.org/stfb2018](https://mouseion.jax.org/stfb2018)

Part of the [Life Sciences Commons](https://mouseion.jax.org/stfb2018), and the [Medicine and Health Sciences Commons](https://mouseion.jax.org/stfb2018)

**Recommended Citation**


This Article is brought to you for free and open access by the Faculty Research at The Mouseion at the JAXlibrary. It has been accepted for inclusion in Faculty Research 2018 by an authorized administrator of The Mouseion at the JAXlibrary. For more information, please contact Douglas.Macbeth@jax.org.
Genome analysis

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

Chris Chatzinakos¹*, Donghyung Lee², Bradley T. Webb¹, Vladimir I. Vladimirov¹, Kenneth S. Kendler¹ and Silviu-Alin Bacanu¹

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

*To whom correspondence should be addressed.

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 Gene-level Joint Analysis of functional SNPs in Cosmopolitan Cohorts (JEPEGMIX) method accommodates cosmopolitan cohorts by using heterogeneous panels. However, JEPEGMIX 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/
Materials and methods

To avoid a mere accumulation of just averagely enriched polygenic information, we competitively adjust \( z^2 \) 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 scan. Due to polygenicity, the expected genome scan \( Z^2 \) statistics, each with 1 degree of freedom (df), has a non-zero background noncentrality parameter \( \lambda \), i.e. \( E(Z^2) = 1 + \lambda \). Thus, by the method of moments, we can estimate \( \lambda^2 = Z^2 - 1 \), where \( Z^2 \) is computed using all measured SNPs in the genome scan. However, given that \( \lambda^2 \geq 0 \), a better estimator is, thus, \( \lambda^2 = \max(Z^2 - 1, 0) \). 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 \( Z^2 \) statistics: \( P^* = 1 - F(Z^2 / \lambda^2) \), where \( F(\cdot / \lambda^2) \) is the cumulative distribution function (cdf) of the non-central \( \chi^2 \) distribution with 1 df and noncentrality parameter \( \lambda^2 \).
2. Transform \( P^* \) into its quantile vector from a central \( \chi^2 \) distribution with 1 df, i.e. \( \chi^2 = F^{-1}(1 - P^*) \).
3. Transform \( \chi^2 \) to a ‘central’ \( Z \)-score: \( Z = \text{sign}(Z) \times \sqrt{\chi^2} \).

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).

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 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.

Table 1. Signals for real datasets

<table>
<thead>
<tr>
<th>Traits</th>
<th>No unique genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCZ</td>
<td>68</td>
</tr>
<tr>
<td>ALZ</td>
<td>34</td>
</tr>
<tr>
<td>AMD</td>
<td>17</td>
</tr>
<tr>
<td>BIP</td>
<td>11</td>
</tr>
<tr>
<td>HDL</td>
<td>79</td>
</tr>
<tr>
<td>LDL</td>
<td>78</td>
</tr>
<tr>
<td>T2D</td>
<td>6</td>
</tr>
<tr>
<td>TG</td>
<td>48</td>
</tr>
<tr>
<td>Smoking</td>
<td>5</td>
</tr>
</tbody>
</table>

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 Illumina 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 PGC2 Schizophrenia ncentrality from PGC2 \( Z^2 \) (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.
Using the Holm \(P\)-value adjustment and both rare (R) and non-

rare (NR) e QTL weights, for the real datasets significant gene sig-
nals 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 abbrevi-
ations 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 co-
horts, we add a competitive version and extend the number of included
(i) eQTLs and (ii) tissues. Unlike existing methods, it also accommod-
est 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 exist-
ing gene expression experiments, the capacity of discriminating causal
tissues will be enhanced by further increases in sample size of such stud-
ies. 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. \textit{et al}. (2015) LD Score regression distinguishes con-
founder from polygenicity in genome-wide association studies. \textit{Nat. Genet.}, 47,
291–295.

Durbin,R.M. \textit{et al}. (2010) A map of human genome variation from


1091–1098.

Gusev,A. \textit{et al}. (2016) Atlas of prostate cancer heritability in European and

7, 10979.


Lee,D. \textit{et al}. (2016) JEPEGMIX: gene-level joint analysis of functional SNPs in


Manolio,T.A. \textit{et al}. (2009) Finding the missing heritability of complex dis-

McCarthy,S. \textit{et al}. (2016) A reference panel of 64,976 haplotypes for genotype

Nica,A.C. and Dermitzakis,E.T. (2008) Using gene expression to investi-
gate the genetic basis of complex disorders. \textit{Hum. Mol. Genet.}, 17,
R129–R134.

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

loci for schizophrenia. \textit{Nat. Genet.}, 45, 1150–1159.


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

Yang,J. \textit{et al}. (2010) Common SNPs explain a large proportion of the herit-