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, Virginia 23298, USA
²The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA

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 (LD) patterns from a relevant reference panel. While most such tools assume homogenous 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 JEPEGMIX, 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.

Contact: chris.chatzinakos@vcuhealth.org

Supplementary information: Supplementary material is available at Bioinformatics online.

1 INTRODUCTION

Gene expression, is believed to have influenced human evolution and play a key role in diseases (Cookson, et al., 2009; Emilsson, et al., 2008; Kudaravalli, et al., 2009). Thus, it is critical for understanding diseases and developing treatments. The importance of gene expression was empirically supported by the enrichment of association signals in SNPs tagging gene expression (Ehret, et al., 2012; Fehrmann, et al., 2011; Fernandez, et al., 2012; Nica and Dermitzakis, 2008; Nicolae, et al., 2010), which are generally known 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 variations associated with a disease. Univariate analysis of GWAS is still the de facto tool for identifying trait/disease-associated genetic variants (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 phenotypic variation (Manolio, et al., 2009). Due to their small effect sizes, these variants are difficult to be 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 METHOD

To avoid a mere accumulation of just averagely enriched polygenic information, we competitively adjust $\chi^2$ statistics for background enrichment (Text S1 Supplementary data (SD)). This is achieved by adjusting the statistic for average non-centrality. Such “centralized” JEPEGMIX statistic we simply denote as competitive (C), while the original statistic we deem as the non-competitive (NC) version. 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 (https://github.com/Chatzinakos/JEPEGMIX2). The annotation file includes cis-eQTL for all tissues available in PREDICTDB.
Using the Holm p-value adjustment and both rare (R) and non-rare (NR) eQTL weights, for the real datasets significant gene signals were found in 9 traits, for which we present heatmaps (Text S6, Figure S6-S23in SD). Table1 shows the number of genes with \(q - value < 0.05\), in the real datasets (for the abbreviations see Table 2 in the SD).

### Table 1. Number of Genes with \(q - value < 0.05\) for the real data (see Table S2 for abbreviations).

<table>
<thead>
<tr>
<th>Trait</th>
<th>No 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>5</td>
</tr>
<tr>
<td>TG</td>
<td>48</td>
</tr>
<tr>
<td>Smoking</td>
<td>5</td>
</tr>
</tbody>
</table>

On a computational node with 4x Intel Xeon 6 core 2.67-GHz processor and 64GB of RAM, the single core schizophrenia JEPEGMIX2 run required slightly under 3.5 hours, including imputation of unmeasured eQTLs. For the remaining datasets the running time was below 3 hours.

## 5 CONCLUSIONS

We propose JEPEGMIX2, an updated software/method for testing the association between (cis-eQTL predicted) gene expression and trait. Unlike existing methods, even for highly enriched GWAS, JEPEGMIX2 competitive version fully controls the Type I error rates at or below nominal levels. To the desirable JEPEGMIX imputing/analyzing capabilities of cosmopolitan cohorts, we, thus, add a competitive version and extend the number of included i) eQTLs and ii) tissues. Unlike the previous version (and all competitors), it also accommodates up weighting of the rare variants and, for high enrichment (above background), it avoids the increased rate of false positives incurred by FDR p-value adjustment by employing a Holm adjustment. Being written in C++, JEPEGMIX2 has very low computational burden. Future versions of the software will be updated to use larger reference panels.

## REFERENCES


Fehrmann, R.S., et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* 2011;7(8):e1002197.


